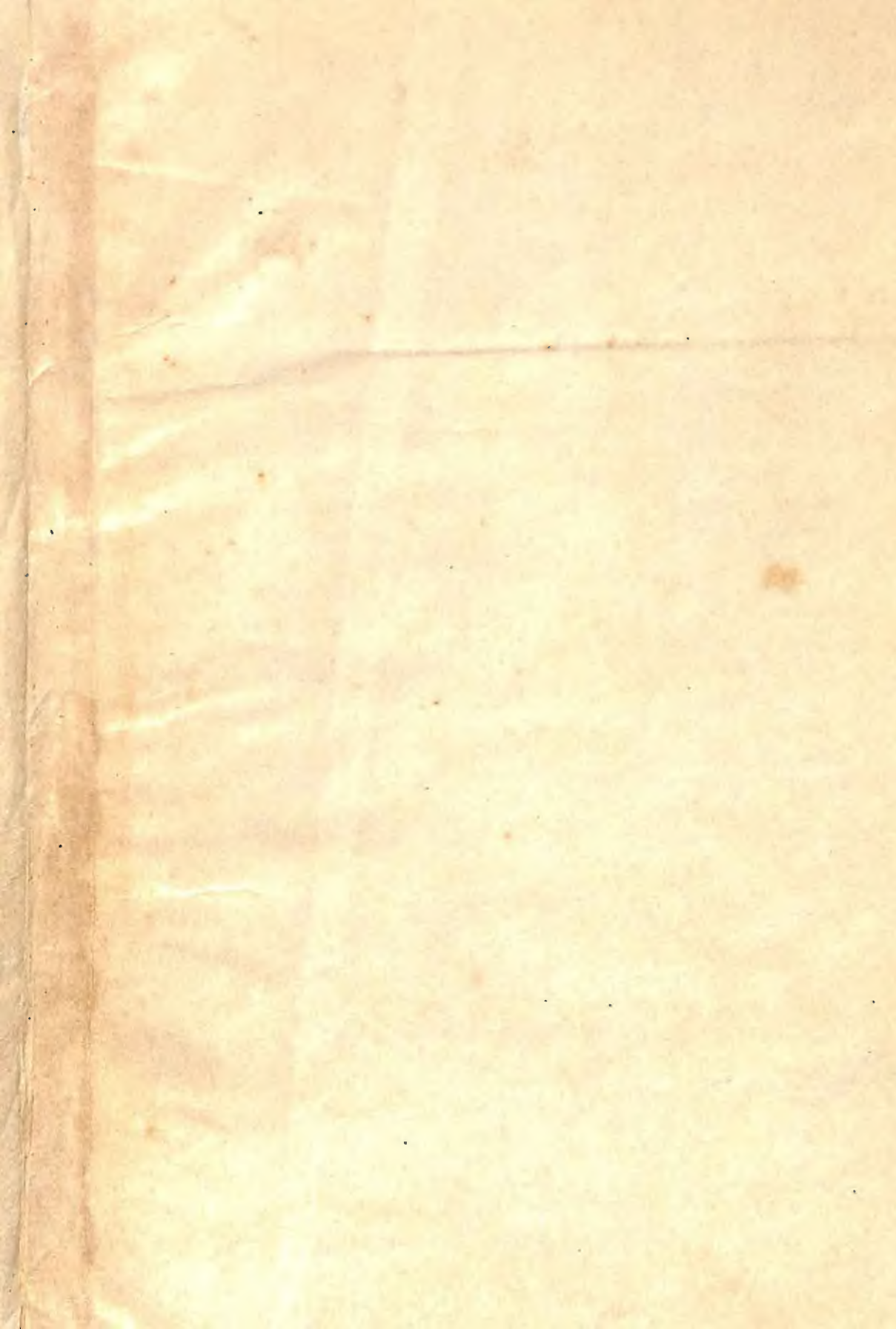


293
1.9.70





JOURNAL OF CONSULTING PSYCHOLOGY

Edited By
Edward S. Bordin
University of Michigan

Associate Editor
Boyd McCandless
University of Iowa

Advisory Editors

John W. Atkinson
University of Michigan
Frank X. Barron
University of California, Berkeley
John M. Butler
University of Chicago
Allen T. Dittmann
National Institute of Mental Health
Seymour Fisher
Baylor University
Harrison G. Gough
University of California, Berkeley
William L. Hays
University of Michigan
Robert B. Malmö
McGill University
Paul E. Meehl
University of Minnesota

Daniel R. Miller
University of Michigan
Alan K. Rosenwald
University of Illinois
Saul Rosenzweig
Washington University
Julian B. Rotter
Ohio State University
Nevitt Sanford
University of California, Berkeley
Seymour B. Sarason
Yale University
Robert R. Sears
Stanford University
Edwin S. Shneidman
Veterans Administration, Los Angeles, California
Read D. Tuddenham
University of California, Berkeley

VOLUME 25
1961

Arthur C. Hoffman, *Production Manager*; Helen Orr, *Promotion Manager*

Published Bimonthly By
The American Psychological Association, Inc.
1333 Sixteenth St., N.W.
Washington 6, D. C.

Printed at
Lancaster, Pennsylvania

Copyright © 1961 by the American Psychological Association, Inc.



Europe Edm. ...
Date 1.8.70 ...
Dated ...
Spec. No. J293 ...

Contents of Volume 25

<i>Adams, Calvin K.</i> See <i>Phares, E. Jerry.</i>	
<i>Adlerstein, Arthur M.</i> See <i>Sperber, Zanwil.</i>	
<i>Affleck, D. C.</i> See <i>Garfield, Sol L.</i>	
<i>Anker, James M.</i> Chronicity of Neuropsychiatric Hospitalization: A Predictive Scale - -	425
<i>Anker, James M., and Walsh, Richard P.</i> Group Psychotherapy, a Special Activity Program, and Group Structure in the Treatment of Chronic Schizophrenics - - - - -	476
<i>Arkoff, Abe.</i> See <i>Bloom, Bernard L.</i>	
<i>Armstrong, Renate G., and Hauck, Paul A.</i> Sexual Identification and the First Figure Drawn	51
<i>Azrin, N. H., Holz, W., and Goldiamond, I.</i> Response Bias in Questionnaire Reports- - -	324
<i>Becker, Wesley C.</i> See <i>Steffy, Richard A.</i>	
<i>Bendig, A. W.</i> Improving the Factorial Purity of Guilford's Restraint and Thoughtfulness Scales - - - - -	462
<i>Berger, Louis, and Liverant, Shephard.</i> Homosexual Prejudice and Perceptual Defense - -	459
<i>Bergs, Lawrence P., and Martin, Barclay.</i> The Effect of Instructional Time Interval and Social Desirability on the Validity of a Forced-Choice Anxiety Scale- - - - -	528
<i>Bledsoe, Joseph C.</i> Sex differences in Mental Health Analysis Scores of Elementary Pupils	364
<i>Block, Jack.</i> Ego Identity, Role Variability, and Adjustment - - - - -	392
<i>Block, Jack.</i> Ego Strength and Conflict Discrimination: A Failure of Replication - - -	551
<i>Bloom, Bernard L., and Arkoff, Abe.</i> Role Playing in Acute and Chronic Schizophrenia -	24
<i>Boomer, Donald S., and Goodrich, D. Wells.</i> Speech Disturbance and Judged Anxiety - -	160
<i>Braen, Bernard B.</i> An Evaluation of the Northwestern Infant Intelligence Test, Test B -	245
<i>Branca, Albert A., and Podolnick, Edward E.</i> Normal, Hypnotically Induced, and Feigned Anxiety as Reflected in and Detected by the MMPI - - - - -	165
<i>Bricklin, Barry.</i> See <i>Piotrowski, Zygmunt A.</i>	
<i>Briggs, Peter F., Wirt, Robert D., and Johnson, Rochelle.</i> An Application of Prediction Tables to the Study of Delinquency - - - - -	46
<i>Burstein, Alvin G.</i> A Note on Time of First Responses in Rorschach Protocols - - - -	549
<i>Carden, Joyce Ann.</i> See <i>Iscove, Ira.</i>	
<i>Cartwright, Desmond S., Robertson, Richard J., Fiske, Donald W., and Kirtner, William L.</i> Length of Therapy in Relation to Outcome and Change in Personal Integration - -	84
<i>Cartwright, Rosalind Dymond.</i> The Effects of Psychotherapy on Self-Consistency: A Replication and Extension - - - - -	376
<i>Cattell, R. B., Knapp, R. R., and Scheier, I. H.</i> Second-Order Personality Factor Structure in the Objective Test Realm - - - - -	345
<i>Chambers, Jay L.</i> Trait Judgment of Photographs and Adjustment of College Students -	433
<i>Chance, June Elizabeth.</i> Independent Training and First Graders' Achievement - - -	149
<i>Coltharp, Frances C.</i> See <i>Stoltz, Robert E.</i>	
<i>Crofts, Irene.</i> See <i>Walters, Richard H.</i>	
<i>Crookes, T. G.</i> Wechsler's Deterioration Ratio in Clinical Practice - - - - -	234
<i>Crowne, Douglas P.</i> See <i>Marlowe, David.</i>	
<i>Cummins, James F.</i> See <i>Dibner, Andrew S.</i>	
<i>Curtin, Mary Ellen.</i> See <i>Estes, Betsy Worth.</i>	
<i>David, Charlotte.</i> See <i>Murstein, Bernard I.</i>	
<i>Dauids, Anthony, DeVault, Spencer, and Talmadge, Max.</i> Anxiety, Pregnancy, and Childbirth Abnormalities - - - - -	74
<i>DeBurger, Robert A.</i> See <i>Estes, Betsy Worth.</i>	
<i>Denny, Charlotte.</i> See <i>Estes, Betsy Worth.</i>	
<i>DeVault, Spencer.</i> See <i>Dauids, Anthony.</i>	

<i>Dibner, Andrew S., and Cummins, James F. Intellectual Functioning in a Group of Normal Octogenarians - - - - -</i>	137
<i>Dickey, Brenda A. Attitudes toward Sex Roles and Feelings of Adequacy in Homosexual Males - - - - -</i>	116
<i>Efron, Herman Y. See Goucher, Elizabeth L.</i>	
<i>Eichman, William J. Replicated Factors on the MMPI with Female NP Patients - - -</i>	55
<i>Epstein, Seymour. Food-Related Responses to Ambiguous Stimuli as a Function of Hunger and Ego Strength - - - - -</i>	463
<i>Estes, Betsy Worth, Curtin, Mary Ellen, DeBurger, Robert A., and Denny, Charlotte. Relationships between 1960 Stanford-Binet, 1937 Stanford-Binet, WISC, Raven, and Draw-a-Man - - - - -</i>	388
<i>Eysenck, H. J. A Note on "Impulse Repression and Emotional Adjustment"- - - - -</i>	362
<i>Farberow, Norman L. See Forer, Bertram R.</i>	
<i>Farkas, Erwin. See Gilberstadt, Harold.</i>	
<i>Feifel, Herman. See Forer, Bertram R.</i>	
<i>Finney, Joseph C. The MMPI as a Measure of Character Structure as Revealed by Factor Analysis- - - - -</i>	327
<i>Fisher, David. See Murstein, Bernard I.</i>	
<i>Fisher, Gary M., Kilman, Beverly A., and Shotwell, Anna M. Comparability of Intelligence Quotients of Mental Defectives on the Weschler Adult Intelligence Scale and the 1960 Revision of the Stanford-Binet - - - - -</i>	192
<i>Fiske, Donald W. See Cartwright, Desmond S.</i>	
<i>Fitzhugh, Kathleen B., Fitzhugh, Loren C., and Reitan, Ralph M. Psychological Deficits in Relation to Acuteness of Brain Dysfunction - - - - -</i>	61
<i>Fitzhugh, Loren C. See Fitzhugh, Kathleen B.</i>	
<i>Forer, Bertram R., Farberow, Norman L., Feifel, Herman, Meyer, Mortimer M., Sommers, Vita S., and Tolman, Ruth S. Clinical Perception of the Therapeutic Transaction - -</i>	93
<i>Furth, Hans G. See Murstein, Bernard I.</i>	
<i>Garfield, Sol L., and Affleck, D. C. Therapists' Judgments Concerning Patients Considered for Psychotherapy - - - - -</i>	505
<i>Gieseeking, Charles F. See Williams, Harold L.</i>	
<i>Gilberstadt, Harold, and Farkas, Erwin. Another Look at MMPI Profile Types in Multiple Sclerosis- - - - -</i>	440
<i>Ginott, Haim G., Lebo, Dell. Play Therapy Limits and Theoretical Orientation - - - -</i>	337
<i>Goldfarb, Allan. Performance under Stress in Relation to Intellectual Control and Self-Acceptance- - - - -</i>	7
<i>Goldiamond, I. See Azrin, N. H.</i>	
<i>Goldstein, Arnold P. See Heller, Kenneth.</i>	
<i>Goodrich, D. Wells. See Boomer, Donald S.</i>	
<i>Goodstein, Leonard D., and Rowley, Vinton N. A Further Study of MMPI Differences between Parents of Disturbed and Nondisturbed Children - - - - -</i>	460
<i>Goodstein, Leonard D. See Heilbrun, Alfred B., Jr.</i>	
<i>Goucher, Elizabeth L., Riggs, Laura E., Efron, Herman Y., Meyers, Rebecca F., and Scanlan, Emily R. Lyons Relationship Scales: A Study of Reliability - - - - -</i>	556
<i>Gouws, David J. Prediction of Relapse for Psychiatric Patients - - - - -</i>	142
<i>Griffith, Richard M., and Taylor, Vivian H. Bender-Gestalt Figure Rotations: A Stimulus Factor - - - - -</i>	89
<i>Guerlin, Wilson H. A Factor Analysis of Geriatric Attitudes - - - - -</i>	39
<i>Guinouard, Donald E. See Rycklak, Joseph F.</i>	
<i>Gynther, Malcome D. The Clinical Utility of "Invalid" MMPI F Scores - - - - -</i>	540
<i>Hamlin, Roy M., and Kinder, Elaine F. Vocabulary Deficit in Brain Operated Schizophrenics- - - - -</i>	239

<i>Hand, Jack, and Reynolds, Herbert H.</i> Suppressing Distortion in Temperament Inventories	180
<i>Hanley, Charles.</i> Social Desirability and Response Bias in the MMPI	13
<i>Hauck, Paul A.</i> See <i>Armstrong, Renate G.</i>	
<i>Haworth, Mary R.</i> Repeat Study with a Projective Film for Children	78
<i>Heath, Helen A.</i> See <i>Korchin, Sheldon J.</i>	
<i>Heilbrun, Alfred B., Jr.</i> The Psychological Significance of the MMPI K Scale in a Normal Population	486
<i>Heilbrun, Alfred B., Jr., and Goodstein, Leonard D.</i> The Relationships between Individually Defined and Group Defined Social Desirability and Performance on the Edwards Personal Preference Schedule	200
<i>Heller, Kenneth, and Goldstein, Arnold P.</i> Client Dependency and Therapist Expectancy as Relationship Maintaining Variables in Psychotherapy	371
<i>Helper, Malcolm M.</i> See <i>Milgram, Norman A.</i>	
<i>Hiler, E. Wesley, and Nesvig, David.</i> Changes in Intellectual Functions of Children in a Psychiatric Hospital	288
<i>Holz, W.</i> See <i>Azrin, N. H.</i>	
<i>Holzberg, Jules D.</i> The Role of the Internship in the Research Training of the Clinical Psychologist	185
<i>Hovey, H. Birnet.</i> An Analysis of Figure Rotations	21
<i>Howe, Edmund S., and Pope, Benjamin.</i> The Dimensionality of Ratings of Therapist Verbal Responses	296
<i>Howe, Edmund S., and Pope, Benjamin.</i> An Empirical Scale of Therapist Verbal Activity Level in the Initial Interview	510
<i>Ingham, J. G., and White, J. M.</i> Sedatives and Suggestibility in Neurotic Patients	182
<i>Iscove, Ira, and Carden, Joyce Ann.</i> Field Dependence, Manifest Anxiety, and Sociometric Status in Children	184
<i>Johnson, Orval G., and Wawerszcek, Frank.</i> Psychologists' Judgments of Physical Handicap from H-T-P Drawings	284
<i>Johnson, Rochelle.</i> See <i>Briggs, Peter F.</i>	
<i>Johnson, Ronald C.</i> See <i>Randolph, Mary H.</i>	
<i>Jones, Austin.</i> Sexual Symbolic Responses in Prepubescent and Pubescent Children	383
<i>Jones, Robert E.</i> Identification in Terms of Personal Constructs: Reconciling a Paradox in Theory	276
<i>Kamano, Dennis K.</i> Self-Satisfaction and Psychological Adjustment in Schizophrenics	492
<i>Kanfer, Frederick, H.</i> See <i>Phillips, Jeanne S.</i>	
<i>Kilman, Beverly A.</i> See <i>Fisher, Gary M.</i>	
<i>Kinder, Elaine F.</i> See <i>Hamlin, Roy M.</i>	
<i>Kirtner, William L.</i> See <i>Cartwright, Desmond S.</i>	
<i>Klahn, James E.</i> See <i>Marks, John B.</i>	
<i>Knapp, R. R.</i> See <i>Cattell, R. B.</i>	
<i>Korchin, Sheldon J., and Heath, Helen A.</i> Somatic Experience in the Anxiety State: Some Sex and Personality Correlates of "Autonomic Feedback"	398
<i>Korman, Maurice.</i> The Concept of Normality: A Reply to Freides	267
<i>Krause, Merton S.</i> Anxiety in Verbal Behavior: An Intercorrelational Study	272
<i>Krause, Merton S., and Pilisuk, Marc.</i> Anxiety in Verbal Behavior: A Validation Study	414
<i>Lacey, Harvey M.</i> See <i>Ross, Alan O.</i>	
<i>LaForge, Rolfe.</i> Objective Estimates of Clinical Judgments	360
<i>Lebo, Dell.</i> See <i>Ginott, Haim G.</i>	
<i>Leibowitz, H. W., and Pishkin, Vladimir.</i> Perceptual Size Constancy in Chronic Schizophrenia	196
<i>Lessing, Elise Elkins.</i> A Note on the Significance of Discrepancies between Goodenough and Binet IQ Scores	456

<i>Levi, Aurelia.</i> Orthopedic Disability as a Factor in Human-Figure Perception - - - -	253
<i>Levine, Gustav.</i> The Effects of Two Verbal Techniques on the Expression of Feelings - -	270
<i>Levinger, George.</i> Social Desirability in the Ratings of Involved and Neutral Judges - -	554
<i>Levitt, Eugene F.</i> See <i>Zuckerman, Marvin.</i>	
<i>Liverant, Shephard.</i> See <i>Berger, Louis.</i>	
<i>Lubin, Ardie.</i> See <i>Williams, Harold L.</i>	
<i>Lubin, Bernard.</i> Judgments of Adjustment from TAT Stories as a Function of Experimentally Altered Sets - - - - -	249
<i>Lubin, Bernard.</i> See <i>Zuckerman, Marvin.</i>	
<i>Lundy, Richard M.</i> See <i>O'Connell, Desmond D.</i>	
<i>Madden, James E.</i> Semantic Differential Rating of Self and of Self-Reported Personal Characteristics - - - - -	183
<i>Marks, John B., and Klahn, James E.</i> Verbal and Perceptual Components in WISC Performance and Their Relation to Social Class - - - - -	273
<i>Marlowe, David, and Crowne, Douglas P.</i> Social Desirability and Response to Perceived Situational Demands - - - - -	109
<i>Martin, Barclay.</i> See <i>Bergs, Lawrence P.</i>	
<i>Martin, Robert C.</i> See <i>Robertson, Malcolm H.</i>	
<i>Matarazzo, Joseph D.</i> See <i>Phillips, Jeanne S.</i>	
<i>Matarazzo, Ruth G.</i> See <i>Phillips, Jeanne S.</i>	
<i>Mednick, Sarnoff A., and Wild, Cynthia.</i> Stimulus Generalization in Brain Damaged Children - - - - -	525
<i>Meyer, Mortimer M.</i> See <i>Forer, Bertram R.</i>	
<i>Milgram, Norman A., and Helper, Malcolm M.</i> The Social Desirability Set in Individual and Grouped Self-Ratings - - - - -	91
<i>Morgan, E.</i> See <i>Rosenberg, B. G.</i>	
<i>Murstein, Bernard I.</i> The Relation of the Famous Sayings Test to Self- and Ideal-Self-Adjustment - - - - -	368
<i>Murstein, Bernard I., David, Charlotte, Fisher, David, and Furth, Hans G.</i> The Scaling of the TAT for Hostility by a Variety of Scaling Methods - - - - -	497
<i>Myers, Rebecca F.</i> See <i>Goucher, Elizabeth L.</i>	
<i>Nesvig, David.</i> See <i>Hiler, E. Wesley.</i>	
<i>Neuringer, Charles.</i> Dichotomous Evaluations in Suicidal Individuals - - - - -	445
<i>O'Connell, Desmond D., and Lundy, Richard M.</i> Level of Aspiration in Hypertensive Cardiac Patients Compared with Nonhypertensive Cardiac Patients with Arteriosclerotic Heart Disease - - - - -	353
<i>Olson, Gordon W.</i> The Influence of Context on the Depression Scale of the MMPI in a Psychotic Population - - - - -	178
<i>Parloff, Morris B.</i> Therapist-Patient Relationships and Outcome of Psychotherapy - -	29
<i>Peterson, Donald R.</i> Behavior Problems of Middle Childhood - - - - -	205
<i>Phares, E. Jerry.</i> TAT Performance as a Function of Anxiety and Coping-Avoiding Behavior - - - - -	257
<i>Phares, E. Jerry, and Adams, Calvin K.</i> The Construct Validity of the Edwards PPS Heterosexuality Scale - - - - -	341
<i>Phillips, Jeanne S., Matarazzo, Ruth G., Matarazzo, Joseph D., Saslow, George, and Kanfer, Frederick H.</i> Relationships between Descriptive Content and Interaction Behavior in Interviews - - - - -	260
<i>Phillips, Leslie.</i> See <i>Zigler, Edward.</i>	
<i>Pilisuk, Marc.</i> See <i>Krause, Merton S.</i>	
<i>Piotrowski, Zygmunt A., and Bricklin, Barry.</i> A Second Validation of a Long-Term Rorschach Prognostic Index for Schizophrenic Patients - - - - -	123
<i>Pishkin, Vladimir.</i> See <i>Leibowitz, H. W.</i>	

<i>Podolnick, Edward E.</i> See <i>Branca, Albert A.</i>	
<i>Pope, Benjamin.</i> See <i>Howe, Edmund S.</i>	
<i>Quast, Wentworth.</i> The Bender Gestalt: A Clinical Study of Children's Records - - - -	405
<i>Rafferty, Janet E.</i> See <i>Tyler, Forrest B.</i>	
<i>Randolph, Mary H., Richardson, Harold, and Johnson, Ronald C.</i> A Comparison of Social and Solitary Male Delinquents - - - - -	293
<i>Raskin, Allen.</i> A Comparison of Acceptors and Resisters of Drug Treatment as an Adjunct to Psychotherapy - - - - -	366
<i>Reisman, John M.</i> An Interpretation of <i>m</i> - - - - -	367
<i>Reitan, Ralph M.</i> See <i>Fitzhugh, Kathleen B.</i>	
<i>Reyher, Joseph, and Shoemaker, Donald.</i> A Comparison between Hypnotically Induced Age Regressions and Waking Stories to TAT Cards: A Preliminary Report - - - - -	409
<i>Reynolds, Herbert H.</i> See <i>Hand, Jack.</i>	
<i>Richardson, Harold.</i> See <i>Randolph, Mary H.</i>	
<i>Riggs, Laura E.</i> See <i>Goucher, Elizabeth L.</i>	
<i>Robertson, Malcolm H., and Martin, Robert C.</i> Sensory Deprivation and Its Relation to Projection - - - - -	274
<i>Robertson, Richard J.</i> See <i>Cartwright, Desmond S.</i>	
<i>Robinson, H. A.</i> See <i>Silverstein, A. B.</i>	
<i>Roos, Philip.</i> Evaluation of Psychotherapy as an Adjunct to Insulin-Coma Therapy - -	450
<i>Rosenberg, B. G., Sutton-Smith, B., and Morgan, E.</i> The Use of Opposite Sex Scales as a Measure of Psychosexual Deviancy - - - - -	221
<i>Ross, Alan O., and Lacey, Harvey M.</i> Characteristics of Terminators and Remainers in Child Guidance Treatment - - - - -	420
<i>Rowley, Vinton V.</i> Analysis of the WISC Performance of Brain Damaged and Emotionally Disturbed Children - - - - -	553
<i>Rowley, Vinton V.</i> See <i>Goodstein, Leonard D.</i>	
<i>Rychlak, Joseph F., and Guinouard, Donald E.</i> Symbolic Interpretation of Rorschach Content - - - - -	370
<i>Saslow, George.</i> See <i>Phillips, Jeanne S.</i>	
<i>Scanlan, Emily R.</i> See <i>Goucher, Elizabeth L.</i>	
<i>Scheier, I. H.,</i> See <i>Cattell, R. B.</i>	
<i>Schonbar, Rosalea Ann.</i> Temporal and Emotional Factors in the Selective Recall of Dreams - - - - -	67
<i>Schonbar, Rosalea A.</i> See <i>Singer, Jerome L.</i>	
<i>Shoemaker, Donald.</i> See <i>Reyher, Joseph.</i>	
<i>Shotwell, Anna M.</i> See <i>Fisher, Gary M.</i>	
<i>Siegman, Aron Wolfe.</i> The Relationship between Future Time Perspective, Time Estimation, and Impulse Control in a Group of Young Offenders and in a Control Group -	470
<i>Silverstein, A. B., and Robinson, H. A.</i> The Representation of Physique in Children's Figure Drawings - - - - -	146
<i>Sindberg, Ronald M.</i> Some Effects of Stimulus Variation on Spiral Aftereffect in Organic and Nonorganic Subjects - - - - -	129
<i>Singer, Jerome L., and Schonbar, Rosalea A.</i> Correlates of Daydreaming: A Dimension of Self-Awareness - - - - -	1
<i>Sommers, Vita S.</i> See <i>Forer, Bertram R.</i>	
<i>Spanner, Marvin.</i> Attribution of Traits and Emotional Health as Factors Associated with the Prediction of Personality Characteristics of Others - - - - -	210
<i>Sperber, Zanwil.</i> Test Anxiety and Performance under Stress - - - - -	226
<i>Sperber, Zanwil, and Adlerstein, Arthur M.</i> The Accuracy of Clinical Psychologists' Estimates of Interviewees' Intelligence - - - - -	521
<i>Spilka, Bernard.</i> See <i>Swickard, Don L.</i>	

<i>Steffy, Richard A., and Becker, Wesley C.</i> Measurement of the Severity of Disorder in Schizophrenia by Means of the Holtzman Inkblot Test - - - - -	555
<i>Stein, Kenneth B.</i> The Effect of Brain Damage upon Speed, Accuracy, and Improvement in Visual Motor Functioning - - - - -	171
<i>Stolz, Robert E., and Coltharp, Frances C.</i> Clinical Judgments and the Draw-a-Person Test	43
<i>Suber, Grace Pennington.</i> Predicting Improvement of Psychiatric Patients from Early Ward Socializing Ratings - - - - -	461
<i>Sutton-Smith, B.</i> See <i>Rosenberg, B. G.</i>	
<i>Swickard, Don L., and Spilka, Bernard.</i> Hostility Expression among Delinquents of Minority and Majority Groups - - - - -	216
<i>Talmadge, Max.</i> See <i>Dauids, Anthony.</i>	
<i>Tatom, Mary Helen.</i> Psychiatric Outpatient Personality Patterns - - - - -	275
<i>Taylor, Vivian H.</i> See <i>Griffith, Richard M.</i>	
<i>Tolman, Ruth S.</i> See <i>Forer, Bertram R.</i>	
<i>Turbiner, Milton.</i> Choice Discrimination in Schizophrenic and Normal Subjects for Positive, Negative, and Neutral Affective Stimuli - - - - -	92
<i>Tyler, Bonnie B.</i> See <i>Tyler, Forrest B.</i>	
<i>Tyler, Forrest B., Tyler, Bonnie B., and Rafferty, Janel E.</i> Need Value and Expectancy Interrelations as Assesed from Motivational Patterns of Parents and Their Children -	304
<i>Van Loan, Malle.</i> See <i>Walters, Richard H.</i>	
<i>Vogel, John L.</i> Authoritarianism in the Therapeutic Relationship - - - - -	102
<i>Wahler, H. J.</i> Response Styles in Clinical and Nonclinical Groups - - - - -	533
<i>Walsh, Richard P.</i> See <i>Anker, James M.</i>	
<i>Walters, Richard H., Van Loan, Malle, and Crofts, Irene.</i> A Study of Reading Disability -	277
<i>Waurzaszek, Frank.</i> See <i>Johnson, Orval G.</i>	
<i>Weiner, Irving B.</i> Cross-Validation of a Rorschach Checklist Associated with Suicidal Tendencies - - - - -	312
<i>Weiner, Irving B.</i> Three Rorschach Scores Indicative of Schizophrenia - - - - -	436
<i>Whitaker, Leighton, Jr.</i> The Use of an Extended Draw-a-Person Test to Identify Homosexual and Effeminate Men - - - - -	482
<i>White, J. M.</i> See <i>Ingham, J. G.</i>	
<i>Wild, Cynthia.</i> See <i>Mednick, Sarnoff A.</i>	
<i>Williams, Harold L., Gieseeking, Charles F., and Lubin, Ardie.</i> Interaction of Brain Injury with Peripheral Vision and Set - - - - -	543
<i>Wirt, Robert D.</i> See <i>Briggs, Peter F.</i>	
<i>Worden, Don Keith.</i> The Intelligence of Boys with Muscular Dystrophy - - - - -	369
<i>Zarlock, Stanley P.</i> Magical Thinking and Associated Psychological Reactions to Blindness	155
<i>Zigler, Edward, and Phillips, Leslie.</i> Case History Data and Psychiatric Diagnosis - - -	458
<i>Zuckerman, Marvin, Levitt, Eugene E., and Lubin, Bernard.</i> Concurrent and Construct Validity of Direct and Indirect Measures of Dependency - - - - -	316

CORRELATES OF DAYDREAMING: A DIMENSION OF SELF-AWARENESS¹

JEROME L. SINGER AND ROSALEA A. SCHONBAR

Teachers College, Columbia University

Despite the fact that most writers on personality theory and psychopathology discuss daydreaming, the highly personal and ephemeral quality of conscious fantasy has posed baffling problems to the investigator who seeks to formulate operational tests of the various theoretical notions in the field. The investigation to be described here represents one phase of a general program of research designed to explore the functional role of daydreaming or fantasy behavior in the organization of personality. Since much of the theory and empirical knowledge concerning daydreaming derives from the observations of individual clinicians with relatively limited subject samples, and under highly specialized conditions (psychoanalysis or examination of psychiatric patients), it was felt desirable to approach the problem from a somewhat different point of view. For one thing, relatively little is known as yet concerning the actual range and variability of daydreaming tendencies in the normal population. There is, furthermore, little systematic knowledge of the relationship of daydreaming tendencies to other personality characteristics or to certain crucial dimensions of behavioral variations in presumably normal individuals. While a variety of studies with thematic apperception type of material have provided useful techniques for scrutinizing patterns of such fantasy needs as achievement and aggression, this research has not primarily been concerned with the more general role of a capacity for daydreaming.

A synthesis of theoretical formulations and some empirical observations by writers such

as Freud, Sullivan, Mead, and Lewin have suggested a view of daydreaming that has served as a basis for some of the tentative hypotheses of the study. The capacity to engage in daydreaming is, to some extent, a learned response which develops differentially as a function of certain patterns of parent-child relationships. Of particular significance in its development appears to be the opportunity for identification with a benign parental figure under circumstances in which intermittent reinforcement for the child's control of overt gratification seeking movements occurs. To some extent, mothers in our society tend to represent inhibition of impulses and also to foster aesthetic interest, while fathers represent action tendencies and the external environment. Closer identification with a mother figure would therefore appear particularly to be related to introspective tendencies.

The mode of translation of checked body movement into a capacity for instituting movement on an imaginal level is difficult to explain; Werner's Sensory-Tonic Theory provides the most specific approach to the problem. With reinforcement both by parental figures and by the general socioeconomic conditions or sociocultural milieu, fantasy or resort to verbal or imaginal means of dealing with delays becomes an increasingly differentiated ability which provides additional benefits, since it frees a person from dependence on the immediate perceptual situation and affords a fluid medium in which trial actions can occur with impunity. In adults, under optimal conditions, a differentiated capacity to engage in daydreaming may make it possible for the individual to increase his awareness of self-other relationships, of his own

¹ This study was supported under Public Health Service NIMH Grant M-2279. The authors are indebted to Vivian McCraven, Judith Antrobus, and John Antrobus, who assisted by acting as raters and in various scoring and computational procedures.

action tendencies seen in time perspective, and it may enhance the possibility of a potentially greater repertory of role relationships through imaginal practice. Pathological extremes in this personality dimension may involve either excessive resort to fantasy with consequent paralysis of fruitful motor exploration of the environment or failure to develop fantasy tendencies (as apparently has occurred in certain institution reared children), with consequent inability to delay motor responses and much self-defeating or destructive motility. The question as to the optimal degree or type of daydreaming remains as yet an unexplored area from the standpoint of empirical research.

To the extent that Rorschach human movement (*M*) responses may represent tendencies to engage in daydreaming (Singer, 1955), support for the notion that daydreaming tendencies are associated with motor inhibition, planfulness, and parental identification has come in a number of studies (King, 1958; Shatin, 1953; Singer & Sugarman, 1955; Singer, Wilensky, & McCraven, 1956). The present investigation represents an effort to move beyond the inferences concerning Rorschach *M* responses to a more direct study of daydreaming and fantasy tendencies. In a somewhat similar effort, Page (1957) recently reported a relationship between a questionnaire derived daydream score and *M*.

HYPOTHESES

The general hypothesis of this study is that subjects (*Ss*) who indicate a greater frequency of daydream behavior are also characterized by greater reported frequency of night dreams, social introversion, and creativity in their spontaneous reports of daydreams or storytelling activity. They are, in addition, more likely to be identified with their mothers (on the basis of measures of assumed similarity of interests); those who report less daydreaming, on the other hand, are expected to show greater evidence of repression or denial of problems and a lesser tendency toward identification with their mothers. The inclusion of a form of manifest anxiety scale (Welsh's *A* scale) in the battery of procedures was carried out with the interest of exploring the possibility of an em-

pirical linkage between daydreaming and anxiety. While it was felt that clinical evidence suggests, generally speaking, a dampening of imaginative behavior during attacks of free-floating anxiety, it was considered likely that the type of behavior reported on the *A* scale might to some extent represent willingness to adopt a self-scrutinizing attitude or to admit complaints, rather than serving as an indicator of gross differences in anxiety.

In effect, then, the conception that is examined empirically in this paper is that one of the dimensions along which people vary involves the tendency or capacity to see themselves in a temporal or spatial perspective and to engage in some form of imaginal living. Operationally, such a tendency or personality style is manifested by relative willingness to respond to questionnaire materials of a personal sort, ability to admit a variety of internal ideational activities, and greater willingness or ability to provide creative thematic material to ambiguously structured stimuli.

SUBJECTS AND PROCEDURE

For the preliminary investigation described here, a group of 44 women, graduate students in education, served as *Ss*. The *Ss* consisted of Negro and white women, married and single, most of whom were teachers. The major group breakdown employed for this study was on the basis of the median score on a questionnaire of daydreaming frequency. No significant differences emerged between the High and Low Daydream groups in age, years of education, marital status, white-Negro ethnic groups, or socioeconomic background.

The following procedures were employed to obtain relevant measures from *Ss* along the hypothesized dimensions:

Daydream Questionnaire. A detailed inventory concerning the patterns of daydreaming and the frequency of occurrence of specific daydreams was developed.² The phase of the inventory employed in the present investigation consisted of a series of 93 specific daydreams. *Ss* were required to indicate on a five-point scale from Very Frequently to Practically Never the relative frequency with which they experienced each daydream. A total score was derived for each *S* based on her self-weighted responses to

² A copy of the questionnaire has been deposited with the American Documentation Institute. Order Document No. 6466 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance \$2.00 for microfilm or \$3.75 for photocopies. Make checks payable to: Chief, Photoduplication Service, Library of Congress.

each item. This Daydream score (ranging theoretically from 93 to 558) served as the basis for dividing Ss into High and Low Daydream groups. A cut at the median score (173) was employed. It should be noted that the internal consistency of the Daydream score was quite high, with Cronbach's alpha yielding a coefficient of .96 for a group of 240 Ss.

Frequency of Night Dreams. Each S kept a log of her night dreams over a period of 1 month. The score employed here was Dream Frequency, the number of separate nights during this period that S remembered at least one reportable dream.

Welsh's Repression Scale—(R). On the basis of an extensive reanalysis of Minnesota Multiphasic Personality Inventory responses, as well as considerable subsequent study, Welsh (1956) developed a scoring scale for MMPI which he terms the R scale. Items on this scale seem best characterized as reflecting for high scorers tendencies toward denial or repression, and for low scorers externalized or acting-out behavior. Most scorable items are answered false for this scale, but Welsh's evidence argues against a simple response set.

Welsh's Anxiety Scale—(A). The A scale, derived similarly by Welsh, consists of items from the MMPI in which "disability of a dysthymic and dysphoric nature" with anxiety is most prominent. According to Welsh's further study of profiles from diagnostic groups, anxiety states fall high on A, but for Ss who score high on both A and R, depression is a primary symptom; those Ss who score high on A and low on R reflect schizoid features.

MMPI Lie Scale. The 15 Lie items from the MMPI were included as an additional measure of denial tendencies and to provide some indication of the extent to which the responses to the daydream questionnaire might be subject to conscious falsification.

Social Introversion—(Si). A scale for social introversion was derived from the MMPI by Drake (1956). Correlations with another measure of introversion were in the .70's for both men and women college students; in addition, the mean for those students engaging in more college activities than the average student showed significantly less introversion than the mean of those participating less than the average amount.

Parental Identification Patterns. As one approach to the issue of similarity to parents, a questionnaire and procedure derived from a study by Oliner (1958) was employed. This questionnaire consists of 44 items dealing with a variety of interests and activity patterns to which Ss indicate their reactions on a four-point scale from "very much like" to "very much dislike." These items were responded to initially by each S for herself, after which instructions called for responding to the questionnaire as "Person I would like to be," and then as the items applied to "Mother" and to "Father." To evaluate the relative perceived similarity of self to mother as against father, a score based on the formula (Self-Father)—(Self-Mother) was derived. A high score on this variable indicates that S reported the difference between her own interests and those of her father to

be greater than the difference between her own interests and those of her mother. A positive correlation was therefore hypothesized between the (S-F)—(S-M) score and degree of fantasy, such that the greater the perceived similarity to mother rather than to father, the greater the fantasy tendency. A score based on the absolute difference between perceived interests of fathers and mothers (F-M) was also employed.

Creativity of Spontaneous Daydream and Storytelling. At the conclusion of the questionnaire, Ss were asked to write an account of an actual daydream and also to make up a spontaneous original story. The daydream and original story were then scored for Creativity, i.e., a measure of the introduction of novel materials, characters, time and space sequences, and emotional vividness. Using a definition of creativity in terms of the above criteria, two examiners independently scored all protocols along a five-point scale for Creativity. Rater reliability for a larger sample of 240 Ss had been evaluated for this variable and was felt to be satisfactory, since in only 11 out of 240 ratings were there differences as great as two points on the five-point scale, and no difference as great as three points. The average of the two raters' scores was employed for the final Creativity score.

Needs Achievement, Self-Aggrandizement, and Affiliation. In addition to scoring the structural characteristics of the story and daydream, some attempt was made to consider the specific thematic content of these materials. Three fantasy needs emerged with enough variability in most of the records to permit a quantitative rating. These were Need Achievement, scored essentially along the lines laid out by McClelland (1958) and Atkinson (1958), Need Self-Aggrandizement (employed here as representing obtaining material possessions or display items, as well as high social status without particular effort or achievement), and Need Affiliation (employed here to include gregariousness, need for social warmth, and sex). It was thought that Need Achievement in particular would relate to degree of daydream activity.

The need scores were rated independently along a five-point scale and raters' results were averaged to give a final score for each S on each need. While these scores could not be considered experimentally independent of the Creativity score, the intercorrelation data below suggest that they cannot be considered merely as reflections of the Creativity score.

Vocabulary Score. To obtain a brief estimate of verbal intelligence, the multiple-choice vocabulary test from the IER Intelligence Scale CAVD was included. This test correlates .50 with a general intelligence factor for a sample of adult males (Thorndike, Norris, & Morrill, 1952), and it provides a simply scored indicator of gross intellectual differences. No specific hypotheses concerning the role of intelligence were formulated for this study, but the Vocabulary score was employed to evaluate the likelihood that particular correlations which emerged might merely represent intelligence differences.

RESULTS AND DISCUSSION

Following dichotomization of the distribution scores for each of the above variables at their medians, tetrachoric r 's were calculated. The matrix of intercorrelations is presented in Table 1.

Inspection of Table 1 reveals general support for the hypotheses in the sense that significant correlations in the predicted direction emerged between the Daydream scores and Dream Recall Frequency, Perceived Similarity to Father minus Perceived Similarity to Mother, Creativity of Spontaneous Daydream and Storytelling material, and Need Achievement. Significant positive correlations also emerged between Daydream score and the A scale, Need Self-Aggrandizement, Need Affiliation and Father-Mother discrepancy. The Repression and Lie scales correlate negatively (at insignificant levels) with Daydream score, while Social Introversion correlates positively as predicted, but at a nonsignificant level. A simple graphic cluster analysis following Tryon (1939) reveals a fairly clear-cut patterning of the variables in this study. Daydreams, Dreams, Social Introversion, Creativ-

ity, Need Achievement, (Father-Mother), and (Self-Father)-(Self-Mother), and the Anxiety scale show a distinct cluster roughly paralleling each other in the extent of intercorrelations in the matrix. The Repression and Lie scales appear to form the negative pole of what appears as a bipolar cluster. Only the Anxiety scale was correlated significantly with Vocabulary; Need Affiliation tends to follow the major cluster with some variations, and Need Self-Aggrandizement does so to a much lesser extent.

Although paralleling the Daydream scale through most of the matrix, the A scale reveals a unique pattern correlating negatively with Need Self-Aggrandizement and Need Affiliation and Vocabulary. Thus, S s who report many problems tend to show fewer fantasy themes dealing with possession and status attainment or need for interpersonal contact. Social Introversion shows a somewhat similar pattern to anxiety, with a particularly high positive correlation emerging with Need Achievement, while a moderate negative correlation is revealed with Need Self-Aggrandizement.

TABLE 1
INTERCORRELATIONS BETWEEN DAYDREAM SCORE AND OTHER VARIABLES

	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Daydreams													
2. Dreams	.36												
3. Repression	-.28	-.22											
3. Anxiety scale	.48	.23	-.71										
5. Lie scale	-.21	.16	.35	-.54									
6. Social Introversion	.25	.22	-.25	.26	-.04								
7. Father-Mother	.31	.00	-.23	.45	-.39	-.04							
8. (Self-Father)-(Self-Mother)	.45	.25	-.52	.31	-.39	.04	-.49						
9. Creativity	.48	.20	-.40	.42	-.48	.15	.53	.21					
10. Need Achievement	.71	.68	-.48	.22	.27	.61	.04	.55	.39				
11. Need Self-Aggrandizement	.41	.20	-.33	-.27	-.41	-.25	.21	.37	.26	.48			
12. Need Affiliation	.48	.28	.06	-.34	.04	.07	.05	-.05	.47	.39	.54		
13. Vocabulary	.07	.08	.21	-.54	-.15	-.18	.00	-.08	.12	.00	-.18	.19	

Easily the dominant variable in the cluster based on size of intercorrelations is Need Achievement. One might infer from this result that much of the achievement need translated into responsiveness to the test situation could account for High Daydream and Night Dream scores, as well as the Creativity and Anxiety scores. This causal type of explanation founders when the high correlation between Achievement and (S-F)-(S-M) is considered, since the latter variable does not lend itself to bias resulting from an achievement or acquiescence set. The concomitant variations of the cluster in question seem accountable on a more complex or subtle basis, therefore.

As a further check on this point, an analysis was made of identification choices of the Ss. As part of the questionnaire, the women in this group were asked to list movie or stage personages, historical figures, and characters from literature whom they emulated or wanted to be like. Analysis of these choices for the High and Low Daydream score groups revealed a significant difference in the "masculinity" or "femininity" of identification figures. The Low Daydream Ss chose significantly more male figures or women engaged in largely masculine pursuits or characterized by traits thought of as predominantly masculine (e.g., Joan of Arc, Elizabeth I of England, Amelia Earhart). The greater "feminine role" or maternal identification of the High Daydream Ss, as well as their high Creativity scores in thematic material and the high Need Achievement scores, suggests support for the suggested relationship between acceptance of inner life or long-range aspirations on the basis of maternal identification.

Evidence supporting the hypothesis relating maternal identification and daydreaming has also emerged for a group of male Ss of comparable background. This male sample did not undergo the same experimental procedures except for the Daydream questionnaire and the self-ratings and will not be reported at length here. Daydream frequency was positively correlated with Self-Father discrepancy ($r = .30$, $N = 64$) and negatively correlated with Self-Mother discrepancy ($r = -.19$, $N = 68$). The results suggest that even for these men, the tendency to perceive oneself as simi-

lar to one's mother and unlike one's father is associated to some extent with reported daydreaming frequency. These results suggest the fruitfulness of exploring daydreaming tendencies and self-awareness variables in terms of their linkages to family constellations and patterns of learning within the family situation or cultural milieu. Only in this way can we hope to move beyond mere classification toward more theoretically derived statements concerning the relationships of a dimension such as "acceptance of inner life" or "self-awareness" to the general framework of personality development.

In conclusion, it appears from these data that there is a general clustering of the variables in a manner suggesting that these women differ along a dimension which might be termed self-awareness. We can, of course, never be sure that High Daydream Ss actually do produce more daydreams and have more conscious achievement-aspirations than Low Daydream Ss. Operationally, we observe only that they accept these phenomena as part of their life-space and report them more readily. It appears likely, however, that the difference in attention and admission to others of these inner experiences may be the psychologically significant phenomenon and that quantitative differences in extent of inner living may be scientifically indeterminable.

SUMMARY

The investigation described here represents one approach to a study of the functional role of daydreaming as a dimension of behavior. It was hypothesized on the basis of various theoretical formulations that Ss who report a high frequency of daydreaming behavior also indicate greater frequency of recall of night dreams, creativity in storytelling, Need Achievement, and, possibly, willingness to admit anxieties or complaints. These High Daydream Ss were expected also to demonstrate greater assumed similarity to their mothers than to their fathers and less evidence of repression or denial (MMPI Lie scale) than low frequency daydreamers. A group of 44 adult female graduate students responded to a variety of questionnaire materials and also reported frequencies of daydreams and night dreams. The test materials

included MMPI Anxiety (A) scale, Repression scale, Lie scale, and Social Introversion (Si), as well as a series of interest items to be filled out by each S for herself, ideal self, and as her mother and her father would have done. Spontaneous storytelling and daydream material were also elicited and scored for Creativity, Need Achievement, Need Affiliation, and Need Self-Aggrandizement. The results supported the general hypothesis, indicating that daydream frequency, night dream recall frequency, thematic creativity, Need Achievement, anxiety, and relatively greater identification with mother than with father intercorrelated positively, while Repression and Lie scales both correlated negatively with the other variables in the cluster. The data suggest that High and Low Daydreamers differ along a dimension which might be termed self-awareness, or acceptance of inner experience.

REFERENCES

- ATKINSON, J. W. (Ed.) *Motives in fantasy, action, and society*. Princeton: Van Nostrand, 1958.
- DRAKE, L. E. Scale O (Social introversion). In G. S. Welsh & W. G. Dahlstrom (Eds.), *Basic readings on the MMPI in psychology and medicine*. Minneapolis: Univer. Minnesota Press, 1956. Pp. 181-183.
- KING, G. F. A theoretical and experimental consideration of the Rorschach human movement response. *Psychol. Monogr.*, 1958, 72(5, Whole No. 458).
- MCCLELLAND, D. C. Methods of measuring human motivation. In J. W. Atkinson (Ed.), *Motives in fantasy, action, and society*. Princeton: Van Nostrand, 1958. Pp. 7-42.
- OLIVER, MARION H. Sex role acceptance and perception of parents. Unpublished doctoral dissertation, Teachers College, Columbia University, 1958.
- PAGE, H. Studies of fantasy: Daydreaming frequency and Rorschach scoring categories. *J. consult. Psychol.*, 1957, 21, 111-114.
- SHATIN, L. Rorschach adjustment and the Thematic Apperception Test. *J. proj. Tech.*, 1953, 17, 92-101.
- SINGER, J. L. Delayed gratification and ego-development: Implications for clinical and experimental research. *J. consult. Psychol.*, 1955, 19, 259-266.
- SINGER, J. L., & SUGARMAN, D. A note on some projected familial attitudes associated with Rorschach movement response. *J. consult. Psychol.*, 1955, 19, 117-119.
- SINGER, J. L., WILENSKY, H., & MCCRAVEN, VIVIAN. Delaying capacity, fantasy, and planning ability: A factorial study of some basic ego functions. *J. consult. Psychol.*, 1956, 20, 375-383.
- THORNDIKE, R. L., NORRIS, R. C., & MORRILL, C. S. General aptitude test battery scores in a regional sample. *USAF Hum. Resources Res. Cent. res. Note*, 1952, No. 52-16.
- TRYON, R. C. *Cluster analysis*. Berkeley: Univer. California Press, 1939.
- WELSH, G. S. Factor dimensions A and R. In G. S. Welsh & W. G. Dahlstrom (Eds.), *Basic readings on the MMPI in psychology and medicine*. Minneapolis: Univer. Minnesota Press, 1956. Pp. 264-281.
- WELSH, G. S., & DAHLSTROM, W. G. (Eds.) *Basic readings on the MMPI in psychology and medicine*. Minneapolis: Univer. Minnesota Press, 1956.

(Received January 25, 1960)

PERFORMANCE UNDER STRESS IN RELATION TO INTELLECTUAL CONTROL AND SELF-ACCEPTANCE¹

ALLAN GOLDFARB

Baltimore City Health Department

In an experimental study, Williams (1947) found a highly significant relationship between Rorschach indices of intellectual control and performance under stress on the Wechsler-Bellevue Digit-Symbol test. His results supported the validity of the Rorschach constructs, indicated that the Rorschach test is a practical instrument for predicting behavior under stress, and implied that efficiency of performance under conditions of stress is mainly a function of intellectual control. This was pointed out by Carlson and Lazarus (1953), who questioned the representativeness of Williams' findings because of his experimental procedures and the conflicting results reported in related studies. They repeated Williams' experiment and did not obtain comparable correlations. These discrepant findings are similar to those contained in a review article on stress by Lazarus, Deese, and Osler (1952).

The present research represents a modification of Williams' experimental procedures in two important respects: in that a real-life stress situation was produced, and the subjects (Ss) had a common motivation to succeed. Measures of self-acceptance were also obtained from the Ss. This personality variable was studied because it is a basic component of psychological theory which indicates that effectiveness of behavior is directly related to self-acceptance (Rogers, 1951; Snygg & Combs, 1949; Symonds, 1951). Of further interest is the congruity between the characteristics of a self-accepting person and the concept of a mature individual having

adequate intellectual control, as depicted by Beck (1945). The prediction was made that acceptance of self would be highly correlated with the Rorschach indices of intellectual control.

METHOD

Subjects

All the pledges ($N = 30$) of a campus fraternity volunteered to participate in this study, which was described as being done to investigate certain important questions facing clinical psychologists. These pledges had been selected by the fraternity from a large number of applicants; they were all highly motivated to become active members. Achieving active status was dependent upon the over-all impression made by each pledge on the fraternity members. This factor was especially crucial during the time this study was being done, since the Ss were in the trial stage of their pledge period. Consequently, an important motivation common to this group of Ss was to make a favorable impression on the fraternity members.

Procedure

In the first phase of the study the Rorschach test was administered individually to each of the pledges, based on Beck's (1944) procedures. This took approximately 1 month.

The Ss then met as a group in a classroom to complete five practice trials on the Wechsler-Bellevue Digit-Symbol test (Wechsler, 1944). During the administration of the sample form of the Digit-Symbol test the experimenter (*E*), in giving instructions to the group, referred to the identical sample form which had been reproduced on the blackboard. Each of five trials of the Digit-Symbol test was completed within 90 seconds, with a 1-minute rest period between each trial. The number of items on the Digit-Symbol test had been increased so that the complete test would not be finished in the allotted time. The Ss then completed the Berger Scale of Expressed Acceptance of Self (Berger, 1952) which yielded the Control Level scores. The Berger scale is self-administering and is composed of 36 items. Selection of these items was made according to the definition of a self-accepting person derived in an earlier study by Sheerer (1949). The respondent rates each item

¹ This paper is based upon a doctoral dissertation completed at the University of Pittsburgh in 1954. The writer is indebted to members of his thesis committee, J. Matthews, A. W. Bendig, H. W. Goodman, and A. D. Lazovik, for their guidance and encouragement.

on a scale from 1 to 5 depending upon how true he feels the item to be in describing himself. To be valid, use of the Berger scale requires that the Ss be unidentified. As it was necessary for the purpose of this study to identify the pledges in order to compare their responses under control and stress conditions, the Berger scale was covertly coded.

The second group session took place the day after the first meeting. Twelve volunteers from the group of assembled pledges were taken to another room, where they completed six additional trials of the Digit-Symbol test. A 3-minute rest period was taken at the end of the third trial, instead of the standard 1-minute interval. The results of the last three trials of this series yielded the Digit-Symbol Control Level scores.

For the final phase of the experiment, the 12 Ss were taken to the psychology department's laboratory in another part of the building to take three more trials of the Digit-Symbol test (Stress Level scores). This procedure of having the Ss complete the Digit-Symbol test under control and stress conditions yielded the behavioral criterion (decrement in performance on the Digit-Symbol test) patterned after Williams' study. In addition, the current criterion measure was designed to incorporate and emphasize psychological stress variables which would be stressful to a pledge to the degree that he was lacking in the characteristics of a self-accepting person as defined by Berger (1952). This was accomplished by exposing the Ss to psychological stress which comprised externally applied pressures as a standard for behavior, maximized their need to deny or distort unacceptable personal characteristics, and indicated that public comparisons would be made of their individual performance results.

In the laboratory, two identical continuous panels had been constructed and placed back-to-back 5 inches apart in the center of a long table, with six positions on each side. Each section of the panel facing the S included a white light and a red light. Extending perpendicularly from the main panel on each side of a given position was a smaller panel. All the panels were 1 foot high. Consequently, when the Ss were completing the experimental tasks, they could not see the progress made by any of the adjacent pledges. All the lights in the laboratory were on; also, two No. 1 photoflood lamps mounted on tripods were placed on the table, one at each end, and beamed at the pledges without causing a direct glare. At one end of the long table in full view of the Ss was the shocking apparatus. This consisted of an inclined panel board which had separate switches for the lights and the electric shock, and was the terminal point for the maze of wires which led from the apparatus to the individual positions.

After the Ss were seated, electrodes were attached to the nonwriting hand of each pledge by E. He then stood at the end of the table near the shocking apparatus, where he could easily be seen by all the pledges. The following instructions were given, patterned after those of Williams:

You are now being observed by a number of psychologists who are taking notes and continuous photographs of all your reactions throughout the remainder of this experiment. [At one end of the table stood a graduate student who operated a portable motion picture camera. Immediately behind each group of three Ss stood a graduate student who served as a judge; the four judges included one female. There was no communication between them while the Ss completed the tests. During the experiment and the rest intervals, they took sham notes of the Ss' behavior. At the end of each trial the judges shifted position, which served to randomize the effect of any particular judge on the S.] All directions are to be followed implicitly. Rest your arm attached to the electrodes on the table and keep it there from now on. You will notice that the electrodes on your arm are now connected to the panel before you. [White light turned on and left on.] The white light that has just gone on indicates that our shock apparatus has been turned on. You are connected to this apparatus. During the following period you may receive a strong electric shock whenever the observers feel that your test performance is not up to our standards. Whenever the red light goes on, you are not meeting our standards and you are in danger of being shocked, like this. [Red light turned on individually for each pledge, and followed by an electric shock. Red light turned off.] Based upon the psychologists' evaluation of your reactions and your tests, each of you will be compared to all the rest of the pledges. You will be compared for personality factors and intelligence. These lists will be posted in your fraternity house 2 weeks from now, so that all of you can see how you compare with the rest of the group. [These results posed a realistic threat to the Ss, who were all highly motivated to make a favorable impression on the fraternity members in order to achieve active status in the fraternity.] Now pick up your pencil and write your name, seat number, and group number in the upper right-hand corner. You will see that this is the same test form you just took downstairs. Your instructions are the same as before. At the signal, "Go," turn the sheet over and work as fast as you can until you are told to stop. Concentrate on your work. Remember, you are being observed and continually photographed. Your work will be compared with the rest of the pledges, and you will be shocked whenever your work falls below our standards. Get set for Trial 1. Go!

Three seconds after the Go signal, the red light was turned on. After a 5-second interval, the electric shock was administered; immediately afterward, the red light was turned off. (The electric shocks were delivered through the electrodes by an electronic interval timer which activated a regulated shock unit. Each pledge could be individually shocked, since the shocking apparatus was connected to a 12-pole position switch. After the shock had been on for 0.4 second, it was automatically turned off. Simultane-

ously, the red light was turned off. The electric shock was then given to the *S* whose position number next appeared on *E*'s list, which had been previously derived by chance selection of the panel position numbers.) Ninety seconds after starting Trial 1, the pledges were told to "Stop." During the 1-minute interval before being instructed to start again, the pledges were again informed of the use to be made of their performance results. The same procedure was repeated in Trials 2 and 3.

After Trial 3 of the Digit-Symbol test was concluded, the electrodes were removed from each *S*'s hand. At that time, to ascertain the sensitivity of the Berger scale to situational influences, the test was taken by the *Ss* under stress conditions (Stress Level scores). They were given instructions similar to those used during the Digit-Symbol testing. As the *Ss* left after finishing the Berger scale, they were cautioned not to return to the room where the other pledges were gathered. The identical experimental procedures were repeated with the remaining pledges, the second group including 12 *Ss*, and the third group consisting of 6 *Ss*.

Shortly after completion of the study, *E* met with the pledges and explained to them the purposes of the research. They were assured of the confidentiality of the data, and that their performance in the study would have no bearing on their status in the fraternity.

RESULTS

The Rorschach records were individually scored according to Beck (1944) and the following three measures of intellectual control were derived: *F*+% for the total record, *F*+% for the color cards alone, and Sum *C*/Total *C*. A high *F*+% is held to indicate a

TABLE 1
SUMMARY OF RORSCHACH PERFORMANCE

Rorschach Category	Experimental group		
	Mean	Range	SD
Sum <i>C</i> /Total <i>C</i>			
Goldfarb	.80	0-1.2	.28
Carlson & Lazarus	.83	.5-1.0	.18
Williams	.88	.5-1.2	.20
<i>F</i> +% Total			
Goldfarb	77.0	52-100	12.28
Carlson & Lazarus	76.8	50-100	13.31
Williams	81.6	70-100	6.59
<i>F</i> +% Color Cards			
Goldfarb	72.5	0-100	20.83
Carlson & Lazarus	70.1	0-100	26.40
Williams	75.5	50-100	9.28

TABLE 2
SUMMARY OF DIGIT-SYMBOL
TEST PERFORMANCE

Digit-Symbol Measure	Experimental Group	
	Mean	SD
1. Control Level		
Goldfarb	88.15	17.16
Carlson & Lazarus	85.1	16.69
2. Stress Level		
Goldfarb	79.63	14.92
Carlson & Lazarus	79.3	15.19
Stress Decrement (1 minus 2)		
Goldfarb	8.52	7.67
Carlson & Lazarus	5.8	8.34
Williams	10.4	5.70
<i>t</i> test		
Goldfarb		5.26*
Carlson & Lazarus		3.41*

* Significant at .01 level.

high degree of intellectual control, while the converse relationship is stated for the Sum *C*/Total *C* measure (Beck, 1944).

Table 1 shows that the *Ss*' scores for the specified Rorschach measures are very consistent with those reported by Williams and by Carlson and Lazarus. This signifies that similar groups of *Ss* were used in all three studies.

The group of pledges reached a plateau of no further improvement on the Digit-Symbol test by Trial 11. This finding corresponded with the results found by Williams, which was not the case in the Carlson and Lazarus study. The Stress Decrement scores in the current study very likely represented a decrement from a level of maximum performance for the pledges. The measure of Stress Decrement was computed by determining the mean number of digits correctly completed for Trials 9, 10, and 11 (Control Level) minus the mean number correctly completed for Trials 12, 13, and 14 (Stress Level). In this study, and in the other two, the magnitude of the Stress Decrement measures was indicative that all *Es* produced comparable stressful conditions by their procedures. These data are contained in Table 2.

TABLE 3
INTERCORRELATIONS AMONG DIGIT-
SYMBOL TEST MEASURES

Digit-Symbol measure	Control Level	Stress Decrement	Corrected Stress Decrement
Stress Decrement			
Goldfarb	.46		
Carlson & Lazarus	.42		
Williams	.05		
Corrected Stress Decrement			
Goldfarb	-.09	.84	
Carlson & Lazarus	-.01	.90	
Improvement under Stress ^a			
Goldfarb	.19	-.25	-.39
Carlson & Lazarus	.09	.33	.32
Stress Maximum ^b			
Goldfarb	.63	.93	
Carlson & Lazarus	.51	.94	
Williams		.93	
Stress Level			
Goldfarb	.86		
Carlson & Lazarus	.87		

^a Last stress trial minus first stress trial.

^b Highest control trial minus lowest stress trial.

Perhaps a more valid criterion measure of experimental stress for the Ss' performance on the Digit-Symbol test is the Stress Maximum score found by subtracting the first stress trial (Trial 12) from the last control trial (Trial 11). It is suggested that Trials 12, 13, and 14 comprise not only decrement in performance under stress, but also include the factor of recovery from stress. The following analysis supports this hypothesis. The difference between the successively larger mean scores made on Trials 12 and 13 was found to be statistically significant beyond the .01 level ($t = 7.56$). However, the mean difference in scores made on Trials 13 and 14 was found to be insignificant ($t = .41$). A possible interpretation of these findings is that Trial 12 represented the primary reaction to the stress situation, while Trials 13 and 14 also reflected the Ss' efforts to recover from and stabilize their reactions to the stress condition. If one were to use the Stress Maximum score as a measure of experimental stress, then the average decrement for this sample was a drop of 15 points. This extremely significant decrement in the Ss' performance on the Digit-Symbol test very likely indicates the period of the greatest influence of the stress factors. Since the Stress Maximum scores correlated very highly with the

Stress Decrement scores ($r = .93$),² as was also found with the Williams and the Carlson and Lazarus studies, they were not correlated with the other measures.

The intercorrelations among the performance test measures are presented in Table 3. A correlation of .36 is required for significance at the .05 level. In this research, as in the one by Carlson and Lazarus, the degree of Stress Decrement on the Digit-Symbol test is correlated with the Control Level. To eliminate the influence of the Control Level of performance, a statistical procedure patterned after that of Carlson and Lazarus (1953, p. 250) was used to derive the Corrected Stress Decrement scores. Although these results were significantly correlated at the .05 level with the measure of Improvement under Stress, the latter scores were also correlated with the various personality indices in order to compare them with the Carlson and Lazarus findings. The Improvement under Stress scores were obtained by subtracting the score for Trial 12 (first trial under stress) from the score for Trial 14 (last trial under stress).

Table 4 indicates that in this study no significant correlations were found between the Rorschach indices of intellectual control and performance under stress, as measured on the Digit-Symbol test. These findings do not support the hypotheses that the Rorschach test

² All coefficients of correlation reported in this study were derived by the Pearson product-moment method.

TABLE 4
CORRELATIONS BETWEEN DIGIT-SYMBOL TEST
MEASURES AND RORSCHACH MEASURES

Digit-Symbol Measure	Sum C	F+%	F+%
	Total C	Total	Color Cards
Stress Decrement			
Goldfarb	-.02	-.14	-.07
Carlson & Lazarus	-.37	.16	.06
Williams	.35	-.61	-.72
Corrected Stress Decrement			
Goldfarb	-.03	-.02	.00
Carlson & Lazarus	-.28	.05	.02
Improvement under Stress			
Goldfarb	.14	.13	-.08
Carlson & Lazarus	-.07	.12	.03

TABLE 5
SUMMARY OF BERGER SCALE PERFORMANCE

Berger Scale Measure	Experimental Group	
	Mean	SD
1. Stress Level	145.03	14.07
2. Control Level	137.87	17.29
	(135.50) ^a	(22.36)
Stress Level minus Control Level ^b	7.16	9.09
<i>t</i> test	4.09*	

^a Data in parentheses obtained by Berger with day college students.

^b Measure of Increase in Self-Acceptance (Stress).

* Significant at .01 level.

can be used to predict behavior under stress, and that reactions to stress are mainly a function of the personality variable of intellectual control. Furthermore, the validity of these Rorschach constructs is not confirmed. The results of the current study are in marked contrast to those found by Williams, but are consistent with the correlations obtained by Carlson and Lazarus.

The mean difference in the Ss' scores on the Berger scale obtained under control and stress conditions was statistically significant at the .01 level. This finding supports the validity of the psychological stress experienced by the Ss; indicates that the Berger scale provides a sensitive measure of self-acceptance; and reveals that the pledges as a group presented themselves as being more self-accepting, i.e., more mature and independent, when they learned that their responses to the Berger scale would be made public. Comparison of the present results derived under control conditions with the findings obtained by Berger with a group of day college students, a sample comparable to the pledges used in this study, indicated consistency of results. Table 5 lists these results.

Analysis of the intercorrelations among the Berger scale measures indicated that Increase in Self-Acceptance (Stress) was significantly correlated at the .05 level with the Control Level ($r = .37$). To eliminate the influence of the Control Level of performance, Carlson and Lazarus' (1953, p. 250) statistical procedures were followed to derive the measure designated the Corrected Increase in Self-Ac-

ceptance (Stress). The latter measure correlated $-.05$ with the Control Level scores, and $.79$ with the Increase in Self-Acceptance (Stress) scores. The correlation between the Control Level scores with the Stress Level scores was $.84$.

The correlations between the Berger scale measures and the Rorschach indices of intellectual control included one statistically significant relationship. The measure of Corrected Increase in Self-Acceptance (Stress) was found to be negatively correlated at the .05 level with the $F + \%$ on the Rorschach color cards ($r = -.38$). This suggests that the Ss with lesser degrees of intellectual control tend to present themselves as being more self-accepting under conditions of stress.

No significant relationships were found between Ss' performance on the Digit-Symbol test and the measures of self-acceptance obtained under control and stress conditions. This finding neither supports the hypothesis that a major personality correlate of behavior under stress is the variable of self-acceptance, nor does it indicate that the Berger scale can be used to predict performance under stressful conditions.

DISCUSSION

No significant relationships were found in the present study between performance under stress and the personality variables of intellectual control and self-acceptance. These Rorschach findings match those of Carlson and Lazarus, but differ markedly from Williams' results.

The possibility that the experimental procedures may have obscured significant relationships merits further study. This concerns the practice of using a single score as a measure of the S's performance under stress which, in turn, is correlated with other scores representing personality variables. A performance score may mask several important components and patterns of behavior. These may not only vary between Ss who achieve identical scores but, if separately correlated with selected components of the personality variables, could conceivably yield significant relationships (see the excellent discussion of this problem by Lazarus et al., 1952).

The increased mean score made by the pledges taking the Berger scale under condi-

tions of stress may be interpreted as a defensive reaction. This is based on the premise that the pledges experienced psychological stress in anticipation that their personal characteristics were not as acceptable to the fraternity members as was indicated on the Berger scale under control conditions. From this viewpoint, rather than being solely a measure of self-acceptance, the increased mean scores include an emotional component of defensive behavior.

No relationship was found between the Berger scale taken under control conditions and the Rorschach test. This finding does not support the initial hypothesis that greater degrees of self-acceptance are associated with increased intellectual control. The significant negative correlation found between the variables of Corrected Increase in Self-Acceptance (Stress) on the Berger scale and $F+\%$ on the color cards of the Rorschach test may be viewed as suggesting that the increased mean score of self-acceptance obtained under conditions of stress primarily reflects defensive behavior, and increased defensiveness by the pledges is related to correspondingly lesser degrees of intellectual control. It should be noted that this significant correlation may have arisen by chance, since it was one of a much larger number of relationships investigated in the study.

SUMMARY

The present study investigated the relationship between performance under stress and the personality variables of intellectual control and self-acceptance. In an attempt to provide a more valid and definitive test of these relationships, the Ss were presented with a realistic stress situation and had a common motivation to succeed. The behavioral criterion was decrement in performance on the Digit-Symbol test. Measures of intellectual control were derived from the Rorschach test, and the Berger scale was used to obtain measures of self acceptance. The following research procedure was observed: (a) the Rorschach test was administered to each of 30 pledges, which took approximately 1 month; (b) the pledges then met as a group and completed five practice trials on the Digit-Symbol test, and took the Berger scale under control

conditions; (c) the following day the pledges met again as a group to complete six more trials on the Digit-Symbol test, the latter three serving as control measures; and (d) they were then immediately taken to the experimental laboratory to complete three more trials on the Digit-Symbol test and to take the Berger scale under conditions of stress.

The major conclusions to be drawn from this study based on the experimental conditions are as follows: (a) support is lacking for use of either the Rorschach test or the Berger scale to predict performance under stress, (b) the personality variables of intellectual control and self-acceptance do not appear to be major correlates of behavior under stress, and (c) confirmation is lacking for the validity of the Rorschach constructs. The Rorschach findings are consistent with those of Carlson and Lazarus, who did not obtain comparable results in a duplicate study of that done by Williams.

REFERENCES

- BECK, S. J. *Rorschach's test*. Vol. I. *Basic processes*. New York: Grune & Stratton, 1944.
- BECK, S. J. *Rorschach's test*. Vol. II. *A variety of personality pictures*. New York: Grune & Stratton, 1945.
- BERGER, E. M. The relation between expressed acceptance of self and expressed acceptance of others. *J. abnorm. soc. Psychol.*, 1952, 47, 778-782.
- CARLSON, V. R., & LAZARUS, R. S. A repetition of Meyer Williams' study of intellectual control under stress and associated Rorschach factors. *J. consult. Psychol.*, 1953, 17, 247-253.
- LAZARUS, R. S., DEESE, J., & OSLER, SONIA. The effects of psychological stress upon performance. *Psychol. Bull.*, 1952, 49, 293-317.
- ROGERS, C. R. *Client-centered therapy*. Boston: Houghton Mifflin, 1951.
- SHEERER, ELIZABETH T. An analysis of the relationship between acceptance of and respect for self and acceptance of and respect for others in ten counseling cases. *J. consult. Psychol.*, 1949, 13, 169-175.
- SNYGG, D., & COMBS, A. W. *Individual behavior: A new frame of reference for psychology*. New York: Harper, 1949.
- SYMMONS, P. M. *The ego and the self*. New York: Appleton-Century-Crofts, 1951.
- WECHSLER, D. *The measurement of adult intelligence*. (3rd ed.) Baltimore: Williams & Wilkins, 1944.
- WILLIAMS, M. An experimental study of intellectual control under stress and associated Rorschach factors. *J. consult. Psychol.*, 1947, 11, 21-29.

(Received December 1, 1959)

SOCIAL DESIRABILITY AND RESPONSE BIAS IN THE MMPI

CHARLES HANLEY
Michigan State University

Two sources of individual differences in responses to self-report inventories stand apart from traditional concepts of personality. The first is the degree to which subjects (Ss) are affected by the "social desirability" of attitudes expressed in inventory items. Scales that measure "defensiveness," "plus-getting," "dissimulation," and "malingering" focus on this factor. The second is seen when Ss are influenced by the form of the answer sheet. Measures of "acquiescence," "response set," and "response bias" are concerned with the effects of response categories. Both kinds of measure, it is hoped, will ultimately be useful in suppressing personality scale variance irrelevant in diagnosis and screening.

A recent study reported by Wiggins (1959) yields important information on a number of scales used for the MMPI. Eleven different measures, nine of which deal with some aspect of social desirability, are compared and found to differ widely in ability to discriminate between protocols of undergraduates instructed to give the socially desirable answer to each MMPI item and protocols obtained under standard instructions. Wiggins draws conclusions from this study that raise general questions regarding past and future work with measures of social desirability. Wiggins distinguishes two approaches to the measurement of test taking defensiveness; these differ in the manner in which scales are constructed and originally validated. A measure built to discriminate Ss given instructions aimed at maximizing defensiveness from Ss taking the inventory under normal conditions has been constructed by the "empirical" method and possesses "empirical" validity. A scale successfully devised to correlate in expected directions with diagnostic scales has been constructed by the "rational" method and has

"rational" validity. The two most effective measures in Wiggins' study are both empirical scales. Wiggins concludes that "empirical methods are the methods of choice" (p. 427) in constructing measures of social desirability. From analysis of correlations between the various measures, Wiggins suggests that earlier studies, presumably those using rational methods, "would be more appropriately considered as studies of response bias" (p. 426). Finally, from reading the paper it is difficult to escape the impression that the empirical method of validation he employs is a close approximation of the real life screening and diagnostic situation.

The purpose of the present paper is to examine: (a) the degree to which effectiveness in his study is consistently related to the empirical-rational distinction as well as to other dimensions he has not considered, (b) whether the influence of response bias in rational measures is as clear as he suggests, and (c) whether empirical validation, when employed with specially instructed and standard groups, is free from defects specific to the procedure. The first point can be clarified by detailed examination of Wiggins' data. The second and third require additional data obtained for the purpose. The analysis that follows is not intended to dispute the potential effectiveness in the real life situation of any specific scale, but rather to consider general procedural questions that bear on the construction of useful measures of the influence of social desirability.

CLASSIFICATION AND EFFECTIVENESS OF MEASURES

Characteristics of measures of defensiveness can be illustrated by referring to eight scales studied by Wiggins. (Three others are omitted

TABLE 1
CONSTRUCTION AND EFFECTIVENESS OF MMPI MEASURES IN WIGGINS (1959) STUDY

Scale	Validation	Item Content	Response Frequencies	Aim	Effectiveness (phi coeff.)
<i>Sd</i>	Empirical	Explicit	Yes	Defensiveness	.721 ^a
<i>Cof</i>	Empirical	Implicit	Yes	Defensiveness	.619
<i>L</i>	None	Explicit	(Guessed)	Defensiveness	.539
<i>Ex</i>	Rational	Explicit	Yes	Def. & Plus-get.	.461
<i>SD^b</i>	Rational	Explicit	No	Defensiveness	.330
<i>K</i>	Empirical	None ^c	Yes	Def. & Plus-get.	.217
<i>Ds</i>	Empirical	Implicit	Yes	Dissimulation	—
<i>B</i>	Rational	None	Yes	Response Bias	—

^a .683 on cross-validation.

^b Labeled *En* in original study.

^c Except for eight items.

in the interest of simplicity; they play a minor part in his analysis.) These are *L* and *K*, both standard MMPI measures, Edwards' *SD* (Fordyce, 1956), *Ex* (Hanley, 1957), *Cof* (Cofer, Chance, & Judson, 1949), *Sd* (Wiggins, 1959), *Ds* (Gough, 1954), and *B* (Fricke, 1957). All but the last are concerned with social desirability.

Wiggins emphasizes the importance of method of validation. To it should be added: (a) use of the results of some type of judgment of item content in determining whether or not to include items in a scale, (b) use of response frequencies to determine inclusion or rejection of items, and (c) the original aim of the scale. Similarities and differences among the eight measures with respect to all of these variables are summarized in Table 1.

Method of Validation. Table 1 indicates that empirical and rational procedures were used in the original validation of most of the scales. The *L* scale, however, was included in the MMPI without a validity study.

Item Content. Item selection for several scales was wholly or partly dependent on *explicit* judgments of item content. The *L* scale, for example, consists of items *written* to allow defensive individuals to claim unrealistically favorable traits. Edwards selected items for his *SD* measure after 10 judges gave socially desirable answers to a pool of *F*, *K*, and Taylor MAS items. Judgments of item content also played an important role in the construction of the *Sd* and *Ex* measures.

The *Cof* and *Ds* scales were derived in part by having certain Ss "fake" roles. These in-

structions seem to involve *implicit* judgment of item content on the part of such Ss. Eight of the 30 *K* items were also chosen on the basis of results of a faking study (Meehl & Hathaway, 1946, p. 543).

Item content was not considered in the derivation of Fricke's *B* measure. Selection without attention to individual item content can be illustrated in the case of the 22 *K* items that constitute the *L6* scale (Meehl & Hathaway, 1946).

In brief, *L6* was derived by an item analysis of the responses of 25 males and 25 females in the psychopathic hospital whose profiles showed an *L* score of $T=60$ or more and who, with the exception of six normal cases, had diagnoses indicating the probability that they should have had abnormal profiles but whose profiles were in reality within the normal range (p. 540).

The item responses of these fifty cases handled separately for males and females were compared to the male and female item frequencies from the general group of males and females that has been used in past scale derivations. In all, 22 items were chosen as a result of this comparison (p. 541).

After these items had been selected, Meehl and Hathaway described them as giving an "over-all impression" of "impunitiveness" (p. 541). That selection on the basis of item content is not the same as interpretation following selection is indicated by the fate of Booklet Item 461, keyed "true" on *Sd*, *Cof*, and *Ex* but "false" on *K*.

Response Frequencies. Several measures were derived wholly or partly by use of response frequencies obtained from groups taking the MMPI. The quotations from Meehl

and Hathaway given in the preceding paragraph indicate the use of frequency data in selecting items for the *L6* scale incorporated into *K*. To obtain the remaining eight *K* items, moreover, response frequencies also played a role. Items in the *Ex* pool were included only if 36–64% of Hathaway's college sample had endorsed them. The *Cof*, *Ds*, and *Sd* scales were constructed in part by comparing frequencies of endorsement obtained from various groups given special and standard instructions.

Response frequencies were not used in the construction of the *SD* measure, although some of its items were drawn from the *K* pool and thus remotely reflect response data. The authors of the *L* scale did not inspect response frequencies for their 15 items but assumed that few honest persons could answer them in the socially desirable direction.

Items for the *B* scale, a measure of response bias, were chosen entirely on the basis of response frequencies, only those endorsed by 40–60% of Hathaway's normative sample being used.

Aim. Most of the measures in Table 1 are oriented toward test taking defensiveness, the tendency to give socially desirable rather than personally relevant answers. *K* and *Ex*, however, also aim at plus-getting, the tendency to be overly critical of oneself. The *Ds* scale is directed at the detection of deliberate plus-getting. The *B* scale, as indicated before, is aimed at response bias rather than defensiveness.

Effectiveness. Wiggins presents extensive data on relative effectiveness of the various scales in discriminating between a sample of 250 college students instructed to give the socially desirable response to each MMPI item and a sample of 190 students taking the inventory under standard conditions. Wiggins determined mean scores separately for men and women. Scales that significantly differentiated records obtained under the two conditions were analyzed to estimate the degree to which this differentiation was accurate. His data, expressed as phi coefficients, are shown in the last column of Table 1. These are based on pooled male and female protocols.

Using the categories in Table 1, we can examine in order the qualities associated with

effectiveness in Wiggins' study. First is validation. The *L* scale does well despite lack of any original validation, although it is by far the shortest scale studied. While the most effective measures are the empirical *Sd* and *Cof* scales, the equally empirical *K* scale is the worst of the lot. Empirical validation, it appears, has no systematic advantage over other types.

A noticeable characteristic of effective measures appears in the item content column of Table 1. The single defensiveness scale not consistently employing attention to item content is *K*, which fares badly.

Data on response frequencies are equally useful. The one defensiveness measure not using such information is *SD*, which is relatively ineffective. Even "guessed" response frequencies, as in the case of *L*, are better than none according to Wiggins' results.

Comparison of the Aim and the Effectiveness columns indicates scales designed to measure defensiveness may have some advantage in Wiggins' study over scales with broader aims. *Ex*, devised to measure both defensiveness and plus-getting, does not fare badly, but *K*, with similar aims, is ineffective. Scales oriented toward behavior other than defensiveness, as in the case of *Ds* and *B*, are completely ineffective, a result that is not unexpected.

In summary, the entries in Table 1 indicate that several characteristics distinguish between measures that were effective and ineffective in Wiggins' investigation. Lack of attention to item content and response frequencies are more clearly associated with ineffectiveness than is the empirical-rational dimension. The advantage for empirical validation is not as systematic as Wiggins' conclusions indicate.

RATIONAL VALIDITY AND RESPONSE BIAS

The superiority of *Sd* and *Cof* is based solely upon empirical validation. For *Ex*, *K*, and *SD*, Wiggins' results reveal superior rational validity, that is, higher correlations with MMPI diagnostic keys. Supporting his preference for the empirical approach is the suggestion that these correlations between defensiveness and diagnostic measures result from response bias. He presents additional

data showing rational measures to be highly correlated with Fricke's *B* scale. These same data, however, indicate that the empirical *K* measure also correlates highly with *B*. Careful consideration of these results is needed.

The measurement of response bias is based entirely on rational validity. Any scoring key with an imbalance of "true" and "false" responses is expected to correlate with any other imbalanced key. When Wiggins reports a correlation of $-.638$ between *K* and *Sc*, for example, it can be attributed to response bias, because all but one of the 30 *K* items are keyed false, and 59 of the 78 *Sc* items are keyed true. The *S* set to answer true should get a high score on *Sc* and a low one on *K*. The reverse holds for the person biased to answer false. Individual differences in defensiveness, however, lead to the same empirical result.

In devising *B*, Fricke (1957) assumed that items of greatest "controversiality" (i.e., items yielding nearly equal numbers of true and false responses) are most susceptible to response bias. The *B* scale is composed of all MMPI items endorsed by 40–60% of Hathaway's normal samples and not appearing on *K*. As Table 1 indicates, *B* is a rational measure constructed entirely from response frequencies.

Securing adequate measures of response bias is made difficult by questions as to the existence of several such biases (Jackson & Messick, 1958; Hanley, 1959). If these problems are set aside, however, another difficulty arises. Should items on the MMPI express undesirable traits more often than desirable ones, the use of response frequencies alone in item selection places the psychologist at the mercy of the manner in which the authors of the inventory worded their items. A scale based entirely on response data may have an excess of items describing undesirable characteristics. If this occurs, response bias and defensiveness are confounded. An individual will tend to obtain a low score, for example, by giving socially desirable answers to items. Correlations between the response bias measure and defensiveness scales then would be partly due to the role of social desirability. This has been suggested regarding correla-

tions between Fricke's OAIS Set T scale (1956) and MMPI measures (Hanley, 1957).

The correlations between *B* and defensiveness measures are inconclusive if it can be shown that *B* is affected by social desirability. Extensive as Wiggins' data are, additional information is needed to settle this question. In the original study of *Ex* (Hanley, 1957), it was recognized that item imbalance might lead to contamination by response bias. For this reason, a second version, *Sx*, containing equal numbers of true and false responses, was described together with data showing that it correlated significantly with several MMPI diagnostic and validating scales. When social desirability was ignored and all *Sx* items keyed true to give a measure of response bias (*AT*), correlations were obtained with several MMPI keys in the predicted direction. *Sx* and *AT* scores, however, were not significantly correlated. From both sets of correlations, it was concluded that many MMPI measures were influenced by both response bias and defensiveness.

B has correlations of $.49$ with *AT* and $-.33$ with *Sx*, computed from the protocols of Hanley's 1957 sample. Both coefficients are significant at the 1% level. These results suggest that *B* is influenced by defensiveness as well as response bias. More direct evidence on this point, however, is obtained from judgments of the social desirability of the items comprising the *B* scale.

Social desirability judgments were available for 25 *B* items from the earlier study (Hanley, 1957). The remaining 38 *B* items, together with three markers that help define the extremes and middle of a nine-point social desirability rating scale, were rated by 26 male and 33 female Michigan State University students in two undergraduate child psychology sections. Social desirability of an item is defined by its median rating. As in the earlier study, items with values of 4 or less were categorized as undesirable and those rated 6 or more desirable, while items with medians between 4 and 6 were treated as neutral.

Of 63 *B* items, 21 were judged undesirable, 32 neutral, and 10 desirable. The scale, it appears, has an imbalance of socially undesirable items.

Another way to demonstrate the imbalance

TABLE 2
INTERNAL CONSISTENCY RELIABILITY
OF SUBSCALES OF *B*

Sex	Empirical Reliability		
	21 Undesirable Items	32 Neutral Items	10 Desirable Items
100 Men	.576	.513	.370
68 Women	.653	.397	.217
Reliability corrected to length of 32 items			
100 Men	.674	.513	.653
68 Women	.741	.397	.470

is to consider reliable variance. If the hypothesis is correct, undesirable and desirable items will be influenced by two sources of systematic variance: social desirability and response bias. Neutral items will be unaffected by social desirability. Neutral items, therefore, should contribute less variance to the *B* scale, provided allowances are made for difference in number of items involved.

To test this hypothesis, the *B* scale was broken into homogeneous subscales of undesirable, neutral, and desirable items. Kuder-Richardson Formula 20 reliabilities computed for the three subscales are shown in Table 2. The *Ss* were 100 males and 68 females, who in 1955 had taken the MMPI in introductory psychology classes at Michigan State University.

Empirical reliabilities are given in the upper half of Table 2. Since the subscales differ markedly in length, these values must be corrected to make comparisons meaningful. We ask, therefore, what reliabilities would be expected if all subscales consisted of 32 items. The entries in the lower half of Table 2, computed by the generalized Spearman-Brown formula (Guilford, 1954, p. 354), answer this question. The desirable and undesirable items have greater internal consistency than the neutral ones, a result in agreement with the hypothesis that these two subscales contain variance associated with social desirability.

Social desirability in *B* can be shown in yet another way. *B* has an internal consistency of .628 in these women and .647 in the men. By keying responses to the 10 desirable items false and scoring all others true, the role of social desirability is increased at the expense of response bias. The internal consistencies of

the revised measure are .673 for the women and .640 for the men, results again demonstrating that *B* is affected by social desirability.

The conclusion that rational validities of certain defensiveness measures should be considered the result of response bias must be strongly qualified whenever it is based on correlations involving *B*. To devise a satisfactory measure of the hypothetical general response bias to inventory items, one should use judgment of content to eliminate an imbalance in socially desirable and socially undesirable items.

EMPIRICAL VALIDATION

The third aim of the present study concerns the extent to which validation of the kind employed by Wiggins risks incorporating variance specific to the procedure. Such variance will be irrelevant to defensiveness as it occurs in diagnostic and screening situations in real life. A clue to one type of such specific variance is given by the fact that the *L* scale is one of the more effective measures in Wiggins' study. A likely source of such specific variance arises in items that are obviously measures of defensiveness and whose keying as such is transparent. An "obvious" item, to borrow Wiener's (1948) term, is one a sophisticated individual will recognize as a trap for defensiveness.¹ The *L* scale is thought to suffer from such obviousness: "At least, one may conclude that the intent to deceive is not often detectable by *L* when the subjects are relatively normal and sophisticated" (Meehl & Hathaway, 1946, p. 538). Obviousness, however, is undesirable only if keying of responses is transparent. When defensive *Ss* identify an item as pertaining to defensiveness, but think that the nonkeyed response is the critical one, the item remains effective. There may exist, therefore, obvious items worthless in practical use because their scoring is transparent, obvious items useful because their scoring is dis-

¹ An item may be "obvious" for some purposes but "subtle" for others. In Wiener's study, for example, the item "I am happy most of the time" is considered a subtle measure of *Pa* but an obvious measure of *Hy*. Obvious defensiveness items, in the same way, may be subtle on other scales.

TABLE 3
PROPORTIONS OF ITEMS RECEIVING DIFFERENT
NUMBERS OF "OBVIOUS" JUDGMENTS
FROM 48 JUDGES

Scale	Number of "Obvious" Judgments Received			
	36-23	21-16	15-9	8-0
<i>L</i>	.40	.33	.20	.07
<i>Sd</i>	.30	.22	.20	.28
<i>Cof</i>	.26	.29	.18	.26
<i>Ex</i>	.19	.42	.27	.12
<i>K</i>	.13	.30	.30	.27
All Items	.25	.24	.26	.24

guised, and items so subtle that scoring is no issue.

In validating procedures of the instructed vs. standard groups variety, obvious-transparent items are as effective as any others in distinguishing between instructed and control Ss. Ss asked to give the socially desirable answer or to fake a role should do so with obvious as well as subtle items. Controls taking the inventory under standard instructions should avoid defensive answers to obvious-transparent items. A scale derived from comparison of the two groups ought to be effective in similar investigations, but many of its items may prove useless in real life applications. For this reason, performance in Wiggins' study alone is an unsatisfactory standard against which to judge various methods of constructing measures of test taking defensiveness.

To determine proportions of obvious items in the empirical and rational scales, 18 male and 30 female students in the sections used 3 weeks earlier for judgments of the *B* scale each selected the 30 to 40 items most obviously measuring defensiveness from a list of the 103 items on *K*, *L*, *Ex*, *Conf*, and *Sd*. Next, they gave the defensive answer to every item they had chosen. From their choices come two kinds of information: obviousness of each item and transparency of its scoring.

Obviousness of Items. Data on obviousness appear in Table 3. The 103 items, several of which occur on more than one measure, are grouped very nearly into quartiles according to the number of obvious judgments received from the 48 judges. The *L* scale clearly is

composed of a relatively large number of obvious items. *K*, on the other hand, is least obvious, a finding that supports the authors of the MMPI in their belief that *K* is the subtler measure.

The other three measures fall between the two extremes. The empirical *Cof* and *Sd* scales have a greater proportion of extremely obvious items than the rational *Ex* measure. If the first two columns of Table 3 are pooled, however, the advantage for *Ex* disappears. At this point, data on judges' agreement on the defensive answer are relevant. Of the 52 items receiving 16 or more judgments of obvious, answers to 8 were confused to the extent that one-fourth or more of the judges disagreed with the majority. One *Cof* and four *Sd* items were subject to such extensive disagreement, but the majority of judges in each case chose the keyed response. One *K* item was so affected, but the majority answer was wrong. Of three *Ex* items disagreed on, only one was answered in the keyed direction by the majority. *K* and *Ex*, it seems, are even less affected by item obviousness than Table 3 indicates.

Transparency of Scoring. Most items judged obvious were answered by the average judge as keyed on individual scales; nevertheless, some were answered in the nonkeyed direction (i.e., judges thought the "honest" answer was the defensive one). For a systematic study of this behavior, responses to all items receiving nine or more judgments of obvious were analyzed—those with fewer are so subtle that transparency is no issue.

The results of this analysis may be expressed by a ratio of incorrect to correct average identifications of the keyed response. With *K*, for example, the ratio is 7/15—7 incorrect and 15 correct identifications out of 22 items receiving 9 or more obvious judgments. Ratios for the other scales are: *Ex* 5/18, *Cof* 1/24, *Sd* 3/26, and *L* 0/15. These data demonstrate that the most effective scales in Wiggins' study tend to be more transparent in scoring than is the case with those he found less useful.

Subtle Items. While the empirical method used by Wiggins and by Cofer et al. is prone to include items undesirable in a measure of defensiveness, the data in Table 3 indicate that it uncovers a fair number of subtle items.

Prominent among those falling into the lowest quartile of obviousness are items whose content includes the words "I like." Some 15 of the 103 items contain this expression, and 10 of these are among the subtlest. An example is the item: "I would like to be a soldier." Five subtle "like" items appear on *Cof*, six on *Sd*, one on *K*, and none on *L* and *Ex*. Lorge (1937) discovered response bias on the like items on the Strong VIB, and Hanley (1959) has presented evidence of a set specific to such items rather than to items in general. A defensiveness measure with many like items can be affected by individual differences in the specific response set. For this reason, subtle like items may be undesirable in a measure of defensiveness.

Despite proneness to a specific set, like items may prove useful in the screening and diagnostic situations. It is possible that *Ss* attempting to portray themselves in an overly favorable manner tend to like almost everything. If this is true, there is no objection to the use of such items, save for the reservation expressed above.

DISCUSSION

The results of Wiggins' study show high empirical validity for the *Sd* and *Cof* measures. Rather than presenting his findings as only a validation and cross-validation of these particular scales, Wiggins has taken the more constructive path of raising general methodological questions that relate to all measures of social desirability. The danger arises that the success of his scale may lead to an uncritical acceptance of the methodological analysis in which he employs the concepts of empirical validation and response bias to account for different efficiencies and correlations in his comprehensive sample of defensiveness measures. By use of his own extensive data, however, it has been possible to show that method of validation is less systematically related to effectiveness than is selection by attention to item content and response frequencies. The relatively ineffective measures in his study, the empirical *K* and the rational *SD* scales, lacked one of these two selection criteria in their construction.

Wiggins properly points to the possible contamination of rational scales by response bias,

but an empirical method has produced one scale, *K*, that is probably so affected. New data on defensiveness in the *B* scale, which he used to measure response bias, indicate that it is affected by social desirability. While construction of rational measures certainly should aim at balancing items to eliminate response-bias contamination (Hanley, 1957), the empirical methods employed by Wiggins and by Cofer et al. appear from data presented in this paper to suffer the limitation of including items that may be too obvious to be useful in real life measurement of test taking defensiveness.

The data that indicate excellent discrimination for the empirical *Sd* and *Cof* measures show good discrimination for the rational *L* and *Ex* scales. The *L* scale is short, and the *Ex* measure was originally presented as a methodological demonstration rather than as a practical instrument (Hanley, 1957). In view of the success of these four scales, it should be emphasized that both empirical and rational methods can work in the contrasted groups' approximation to the real life situation. (Correlations among these measures in Wiggins' sample of control men demonstrate that *Sd* and *Cof* do not form a pair clearly distinguished from the other measures. *Cof* and *Ex*, for example, correlate more highly than *Cof* and *Sd*, despite a 14-item overlap in the latter two scales.)

Validation by contrasted groups, however, remains only an approximation to screening and diagnostic performance. For this reason, Wiggins' results do not foreclose the possibility that a seemingly ineffective scale may be useful in actual practice. For any defensiveness measure to aid in screening and diagnosis, moreover, one condition must be met: if a linear regression model is used, the defensiveness scale must correlate with the diagnostic measure, that is, a scale cannot suppress irrelevant variance in a predictor unless it correlates with the predictor. Rational scales seem more to meet this requirement than do empirical measures, save for *K*. The correlation of $-.091$ between *Sd* and *Sc* reported by Wiggins, for example, means that *Sd* cannot suppress variance in *Sc* that is associated with defensiveness. While it is possible to hold that the role of defensiveness in responses to in-

ventories is seriously exaggerated—that the low correlation is the outcome of honest answers to Sc items—so many psychologists have assumed defensiveness operates to reduce the effectiveness of inventories that this plausible but radical hypothesis needs verification in studies of protocols obtained from patients and controls in screening and diagnostic settings. Until then, no method can rightly be termed “the method of choice.”

The derivation of useful suppressor scales is not the only concern in studies of social desirability; there ought to be some explanation of failure and success. The methodological considerations raised by Wiggins and by the present paper are important for this reason. Wiggins indicates what he believes is the important dimension; the present paper presents alternatives that fit his results. The final resolution of these differences, however, rests on studies as extensive as that of Wiggins but conducted with actual rather than simulated protocols. The methodological analyses indicate dimensions that should be explored in such a study.

SUMMARY

MMPI scales related to social desirability differ in use of response frequencies and attention to item content in selection of individual items. A reinterpretation of data from an extensive study by Wiggins (1959) indicates that scales using both response frequencies and judged item content in their construction are superior at discriminating controls from subjects instructed to respond to the MMPI in a socially desirable manner. Whether scales were originally validated by the “empirical” or the “rational” method is less systematically related to their effectiveness.

The role of response bias in producing correlations between rational scales and MMPI diagnostic measures is unclear because of the

difficulty of obtaining a pure measure of response bias. The *B* scale employed by Wiggins to measure response bias is also influenced by social desirability.

Derivation by contrasted groups, the empirical method used by Wiggins, suffers from the fact that it includes many items that are obvious measures of defensiveness and whose scoring is transparent.

Preference for empirical or for rational procedures should await studies of their effectiveness in real life diagnostic and screening situations.

REFERENCES

- COFER, C. N., CHANCE, JUNE, & JUDSON, A. J. A study of malingering on the Minnesota Multiphasic Personality Inventory. *J. Psychol.*, 1949, 27, 491-499.
- FORDYCE, W. E. Social desirability in the MMPI. *J. consult. Psychol.*, 1956, 20, 171-175.
- FRICKE, B. G. Response set as a suppressor variable in the OAS and MMPI. *J. consult. Psychol.*, 1956, 20, 161-169.
- FRICKE, B. G. A response bias (*B*) scale for the MMPI. *J. counsel. Psychol.*, 1957, 4, 149-153.
- GOUGH, H. G. Some common misconceptions about neuroticism. *J. consult. Psychol.*, 1954, 18, 287-292.
- GUILFORD, J. P. *Psychometric methods*. (2nd ed.) New York: McGraw-Hill, 1954.
- HANLEY, C. Deriving a measure of test-taking defensiveness. *J. consult. Psychol.*, 1957, 21, 391-397.
- HANLEY, C. Responses to the wording of personality test items. *J. consult. Psychol.*, 1959, 23, 261-265.
- LORGE, I. Gen-like: Halo or reality. *Psychol. Bull.*, 1937, 34, 545-546.
- MEEHL, P. E., & HATHAWAY, S. R. The *K* factor as a suppressor variable in the Minnesota Multiphasic Personality Inventory. *J. appl. Psychol.*, 1946, 30, 525-564.
- WIENER, D. N. Subtle and obvious keys for the Minnesota Multiphasic Personality Inventory. *J. consult. Psychol.*, 1948, 12, 164-170.
- WIGGINS, J. S. Interrelationships among MMPI measures of dissimulation under standard and social desirability instructions. *J. consult. Psychol.*, 1959, 23, 419-427.

(Received December 7, 1959)

AN ANALYSIS OF FIGURE ROTATIONS

H. BIRNET HOVEY¹

Veterans Administration Hospital, Salt Lake City, Utah

When a patient produces a rotation or reversal for a stimulus figure during a psychological test, this is interpreted as a sign of brain damage. For instance, higher weighted scores for brain damage are assigned for rotations than for any other kinds of errors in the Graham-Kendall (1946) Memory-for-Designs Test (G-K). During routine administration of that test at this hospital, it has been found that an occasional intelligent patient produces a rotation without making any other scored error. A correctly reproduced figure, although rotated or reversed, seems on inspection to indicate better cerebration than a reproduction in which the gestalt is changed or other kinds of errors are made for the same figure. So, since brain damaged persons produce rotations to a significantly greater extent than do others, it was hypothesized that such behavior might have significance also in another way. Could it represent an element of bluffing, a hysterical maneuver, a form of role-playing to overdemonstrate brain injury, contrariness, lack of interest, distraction, transient physiological disturbances in the brain? Inspection of a few handy cases with rotation suggested the last as a likely possibility.

PROCEDURE

Of the patients who had had electroencephalograms (EEGs) all the G-K protocols at this hospital

¹ Kenneth A. Kooi, now at the University of Michigan Medical School, turned over to the author comprehensive EEG data which were used in this study. Reed S. Boswell, research psychologist at the Veterans Administration Hospital in Salt Lake City, rechecked G-K scores based on rotations made by the original examiners and gave useful suggestions. Leonard W. Jarcho, Chief of the Neurology Service of this hospital, and Chairman of the Division of Neurology of the University of Utah College of Medicine, critically evaluated the results and interpretations and made helpful suggestions for the manuscript.

containing rotations or reversals were collected to find out if a patterning of some kind might emerge when these protocols were compared as a group with other data in the clinical files. Almost at once the observation was made that there may be a loading for epilepsy. There were 42 protocols with rotations or reversals among a total of 338 protocols for patients who had also had EEGs.

Between 25-50%, varying from month to month, of all patients admitted to this hospital since it opened in 1952 have received EEGs. Cases have been referred primarily for EEGs when shock therapy was contemplated, when epilepsy or brain damage was considered a possibility or was known, when patients were regarded as alcoholics, as elderly psychotics, as special diagnostic problems, etc. When a patient was referred for both an EEG and a psychological evaluation, both referrals were usually made at about the same time.

All the protocols were separated into two major groupings based on EEG summary impressions of the electroencephalographers. The criterion for the first grouping was "normal" or "within normal limits." This contained 129 cases and was set aside. The second major grouping, containing 209 cases, was broken down into two groups for the analysis. All had "abnormal" or "borderline abnormal" records such as "generalized slow patterns," or "slow alpha," etc. When, in addition, mention was made of the presence of transient episodic disruptions of the abnormal background or usual patterns, the case was placed in Group I. Notations for these cases were such as "paroxysmal formations," or "scattered sharp formations," or "spikes," or "occasional delta," etc. This group contained 82 cases. Group II was composed of the 127 cases with abnormal EEGs but without any notations of transient disturbances.

RESULTS

When Group I was compared with Group II for frequency of rotations, the χ^2 was 19.67, which is beyond the .001 level of confidence. The r_{tet} was .57. When Group I was compared with the normal group or a combination of this and Group II, the relationship was higher, as there were only three cases in the normal grouping with rotation. Group I contained 28 such cases.



When the cases were rearranged according to diagnoses and without regard to EEG findings, relationships turned out to be appreciably weaker. A group of 77 patients with the diagnosis of epilepsy was compared with a group of 136 with the diagnosis of brain damage, for incidence of rotation. The χ^2 fell down to 6.47, which reaches only the .02 level of confidence, and the r_{tet} was only .33.

DISCUSSION

One explanation for the drop in relationships when comparisons were made according to diagnoses, is that often an "organic" has seizures of some kind without any mention of this fact in the diagnosis, perhaps because the seizures are regarded by a diagnostician as a minor symptom. Another is that seizures may not have been observed by trained professional personnel. Still another is that a patient may never have had a recognized seizure. At any rate, the data demonstrate that rotations or reversals of G-K figures are associated with transient, episodic EEG disturbances, and they suggest that rotations may be a sign of epilepsy or a potential epileptic condition, possibly a subclinical manifestation.

An interpretation of rotation could be as follows: the subject correctly perceives the stimulus figure, but by the time he starts to draw the reproduction, some kind of transient physiological dysfunction in the brain has occurred, altering his memory of it. There were five rotation cases among the whole series, each with a total error score of 3 on the G-K, these scores being based on one rotation and with no other scorable errors. This would suggest that each person correctly perceived all the figures, but had a crucial interference of vigilance while negotiating one of them.

Frequently an epileptic patient, while doing the G-K test, might announce that he had forgotten the design. Alternatively, he might reproduce it incorrectly and then do it again correctly without prompting. So far tabulations for these behaviors have not been made here. When confronted with an incorrect reproduction after the test was over, an epileptic subject frequently recognized the error and explained it in terms of momentary confusion,

or would say that he had been distracted by thinking of something else, or was not paying enough attention. Prior to the confrontation, most subjects when asked could correctly point out reproductions which were inaccurate, suggesting an awareness of transient confusion. Now and then a subject might make a rotation or a marked error in an otherwise accurate record, and when asked after the test to find the error, would not only locate it but reproduce it correctly without reviewing the stimulus card.

Somewhat comparable phenomena in examinations with the Wechsler Adult Intelligence Scale have already been reported (Hovey & Kooi, 1955; Kooi & Hovey, 1957). MMPIs administered to most of the same subjects in those studies also tended to produce characteristic profiles for epileptics (Hovey, Kooi, & Thomas, 1959).

The majority of the diagnosed epileptic group and also the majority of the group with episodic EEG features had known brain damage. Only 2 of the 42 rotation cases had neither an abnormal EEG nor a diagnosis implying organicity.

Chorost, Spivack, & Levine (1960) report that rotation of Bender-Gestalt figures by children was slightly associated with EEG abnormality but not enough for predictive purposes. The difference between my results and theirs could be explained by the finding that children generally have less ability than do adults to execute the drawing of designs (Pascal, 1951, pp. 23, 42). Therefore control groups of children might be expected to have a relatively high incidence of rotation. The adult groups used in the present study contained much smaller proportions of rotation than did theirs, approaching the vanishing point for subjects with normal EEGs. Furthermore, the present project used 45° instead of their 30° criterion for rotation. However, direct comparisons between the two studies cannot be made since standard administration of the Bender figures permits continuous reference to the stimulus figures, whereas a memory factor is involved in the G-K administration. The current observations are consistent with the prevailing opinion that rotation is associated with organic disease of the adult brain.

SUMMARY

The performance on a design reproduction test of a group of patients with transient episodic EEG features was compared with a group having abnormal EEGs but without observed episodic features. The episodic group made rotations to a significantly greater extent.

REFERENCES

- CHOROST, S. B., SPIVACK, G., & LEVINE, M. Bender-Gestalt rotations and EEG abnormalities in children. *J. consult. Psychol.*, 1960, 23, 559.
- GRAHAM, F. K., & KENDALL, B. S. Performance of brain-damaged cases on a Memory-for-Designs Test. *J. abnorm. soc. Psychol.*, 1946, 41, 303-314.
- HOVEY, H. B., & KOOR, K. A. Transient disturbances of thought processes and epilepsy. *AMA Arch. Neurol. Psychiat.*, 1955, 74, 287-291.
- HOVEY, H. B., KOOR, K. A., & THOMAS, M. H. MMPI profiles of epileptics. *J. consult. Psychol.*, 1959, 23, 155-159.
- KOOR, K. A., & HOVEY, H. B. Alterations in mental function and paroxysmal cerebral activity. *AMA Arch. Neurol. Psychiat.*, 1957, 78, 264-271.
- PASCAL, G. R. *The Bender-Gestalt test*. New York: Grune & Stratton, 1951.

(Received December 19, 1959)

ROLE PLAYING IN ACUTE AND CHRONIC SCHIZOPHRENIA¹

BERNARD L. BLOOM

Hawaii State Hospital

AND

ABE ARKOFF

University of Hawaii

Recent research has suggested that role playing or empathic ability is related to general adjustment. Working with college populations, a number of investigators (Dymond, 1948, 1949, 1950; Lindgren & Robinson, 1953; McClelland, 1951; Norman & Ainsworth, 1954) have found that better adjusted students play roles and empathize with greater facility than those who are less well adjusted. By logical extension it might be assumed that "normals" generally are more skilled in this function than neurotics and psychotics. Some very recent studies, however, have shown that certain schizophrenic groups have considerable role playing skill.

Jackson and Carr (1955) reported, for example, that their normal controls demonstrated greater empathic ability than schizophrenic patients; their schizophrenic sample, however, was quite heterogeneous, some patients consistently demonstrating more empathic ability than a number of controls. When Helfand (1956) compared the empathic ability of four groups—normals, nonpsychotic patients (tuberculous), and chronic and privileged schizophrenics—privileged schizophrenics proved to be superior to all others including normals. Some related information was produced by Grayson and Olinger (1957), who found that when asked to simulate "normalcy," most of their psychiatric patients (largely schizophrenics) were able to improve their test performance and that improvement was related to early discharge from the hospital.

¹ This investigation was supported by a research grant (M-1529) from the National Institute of Mental Health, National Institutes of Health, United States Public Health Service.

Presented, in part, at the meetings of the Western Psychological Association, San Diego, April 17, 1959.

The present study attempted to throw further light on role playing in schizophrenia. On the basis of previous research, the following hypotheses were formulated:

1. Acutely ill schizophrenics are better able to play the normal role than chronically ill schizophrenics.

2. Whether acutely or chronically ill, schizophrenics who subsequently improve are better able to play the normal role than those who do not.

METHOD

Subjects

The subjects (Ss) of the study were 54 hospitalized women diagnosed as either acute or chronic schizophrenics. Each of these two groups was further divided into fast and slow improvement subgroups. Half of the total sample was Caucasian. The remainder was Oriental or part-Hawaiian with the majority being Japanese.

The acutely ill group was made up of 25 patients, none of whom had a history of prior hospitalization beyond that associated with usual commitment procedures. In addition, suddenness of onset in a previously compensated personality structure was also used as a criterion which determined inclusion in this group. Within the acute group, 12 patients were considered to be in the fast improvement subgroup and 13 in the slow improvement one. This classification was based on an evaluation of the hospital course over the 6 months following the testing of the last case. All of the patients of the fast improvement subgroup had been discharged as improved or recovered. Their range of hospitalization was from 1 to 5 months with an average of 2.6 months. In the slow improvement subgroup, seven patients had been discharged; six remained in the hospital. Length of hospital stay for these Ss ranged from 6 to 20 months with an average of 10.9 months. In general, recovery in the slow improvement subgroup was not only less rapid, but it was also qualitatively less striking.

The chronic group was made up of 29 patients who had either been continuously hospitalized for 2 or more years or had been admitted at least twice and had a history of long-standing schizophrenic adjust-

ment. Within this group, 13 patients were placed in the fast improvement subgroup and 16 in the slow one. All of the patients of the fast improvement subgroup had been discharged as improved. Their length of hospitalization (including all former periods) ranged from 8 to 65 months with an average of 29.2 months. In the slow subgroup, no patients had been discharged or seemed ready for even preliminary consideration for discharge. Length of hospital stay for these Ss ran from 12 to 180 months with an average of 56.3 months.

Prior to inclusion in the study, patients under consideration were administered the Vocabulary subtest of the Wechsler-Bellevue, Form I, and only testable patients with a weighted score of seven or higher were used. None of the groups or subgroups differed significantly from others on vocabulary score or amount of formal education. The average number of school grades completed was 10.5. The chronic groups averaged 34 years old; the acute groups' average age was 30.5.

Procedure

In studies of role playing and empathy, the S is usually asked to predict the response of another person who is in some way known to the S. This procedure has been criticized by a number of investigators. Hastorf and Bender (1952) emphasized that projection rather than empathy may account for part of the prediction of another person's responses. Lindgren and Robinson (1953) pointed out that instead of truly empathizing, S may respond in terms of a stereotype; and Helfand (1956) indeed found that his normals tended to rely on a conventional frame of reference although this was not true for his schizophrenic groups, who were deficient in such a reference.

Some investigators have explicitly undertaken to assess their Ss' awareness of normative data rather than their awareness of a specified criterion individual. Indeed, Crow (1959) suggested that when judges are asked to predict personality characteristics of criterion Ss, their judgments are more accurate if they are based upon stereotypes than if they are based on specific information about each criterion S. In Crow's study, a variety of judges (student nurses, medical students, psychiatric residents, and others) were asked to estimate the age, intelligence, vocabulary level, and personality characteristics of 10 medical patients, based upon their seeing a 6-minute sound movie of each patient being interviewed by a physician. In addition, the judges were asked to make estimations for the "average patient." On the basis of these two kinds of judgments, it was possible to compute a stereotype accuracy score (subtracting a judge's estimation for the average patient from each of the criterion scores) and an individual accuracy score (subtracting a judge's estimation for each patient from that patient's criterion score). Crow found that stereotype accuracy was clearly more accurate for estimation of personality characteristics; that is, the judges would have been more accurate if they had given their estimation for the average patient

each time instead of making an individual prediction for each patient.

The procedure used in the present investigation was similar to Crow's and to the one used by Grayson and Olinger (1957) in that awareness of normative data was measured, that is, "abnormal" Ss were asked to play or simulate the "normal" role. Each S was tested in two sessions, the first session in the morning and the second in the afternoon of the same day. In the case of the acutely ill group, the testing was accomplished within 1 week after hospitalization. In the first session, the Rorschach and the Sc scale of the Minnesota Multiphasic Personality Inventory were given under standard instructions. In the second session, these two tests were administered again with special role playing instructions to the S to respond in the way that a "typical, average, ordinary" person would. The instructions were repeated with each Rorschach card and wherever it seemed indicated on the MMPI Sc. The word normal itself was not used because preliminary investigation revealed that this term provoked negative reactions on the part of some patients.

Each Rorschach protocol was scored for the "principal indicators of schizophrenic disorganization" described by Schafer (1948). These indicators include low form level, use of pure color, sex responses, sudden changes, irregular sequence of locations, and various types of deviant verbalizations. Every indicator was given a score of one point each time it appeared except that $F+%$ between 50 and 59 was given a score of 1, and under 50 a score of 2. For each protocol, the total number of indicators was divided by the number of responses to yield a schizophrenic disorganization quotient which took into account the productivity of the S. Statistical analysis of the Rorschach results was based on this quotient. The MMPI Sc was scored in the usual manner.

RESULTS

The Rorschach schizophrenic disorganization indices and the MMPI Sc scores for the various groups under the two experimental conditions are presented in Table 1. On both the Rorschach and MMPI Sc, high scores were regarded as evidence of schizophrenia and reduced scores under role playing instructions were considered to be evidence of ability to play the normal role.

On the Rorschach, all experimental groups showed some reduction in sign of schizophrenic disorganization under role taking conditions. The acute group significantly reduced its quotients, demonstrating a decreased schizophrenic disorganization when playing the normal role. The fast improvement group similarly reduced its quotients. However, the degree of reduction of the signs

TABLE 1
INDICES OF SCHIZOPHRENIC DISORGANIZATION ON THE RORSCHACH TEST AND MMPI *Sc* SCALE
UNDER STANDARD AND ROLE PLAYING INSTRUCTIONS

	Rorschach Test						MMPI Sc Scale			
	Disorganization Indicators		Disorganization Quotient		<i>r</i>	<i>t</i>			<i>r</i>	<i>t</i>
	Mean	SD	Mean	SD			Mean	SD		
Acutely Ill (<i>N</i> = 25)										
Standard	11.36	14.00	.76	.90	.42	2.75**	32.96	17.52	.34	.53
Role Playing	5.16	5.48	.32	.36			30.72	18.40		
Chronically Ill (<i>N</i> = 29)										
Standard	8.93	9.17	.76	.82	.85	.67	22.28	11.90	.44	1.00
Role Playing	7.34	8.28	.70	.78			25.07	15.48		
Fast Improvement (<i>N</i> = 25)										
Standard	8.08	7.36	.67	.66	.15	2.43*	27.48	16.64	.55	.96
Role Playing	4.00	5.04	.33	.33			24.44	16.28		
Slow Improvement (<i>N</i> = 29)										
Standard	11.76	14.21	.84	.99	.76	1.25	27.00	14.83	.30	.96
Role Playing	8.34	8.14	.69	.80			30.48	17.34		

* Significant at .02 level.

** Significant at .01 level.

* Significant at .02 level.

** Significant at .01 level.

of schizophrenic disorganization in the chronic group and in the slow improvement group was slight and did not achieve significance.

None of the MMPI *Sc* results achieved statistical significance. It is interesting to note, however, that both the acutely ill and fast improvement groups appeared to reduce their MMPI *Sc* scores in assuming the normal role, while both the chronically ill and slow improvement groups appeared to increase their scores on the same test under role taking conditions.

The statistical tests of the two experimental hypotheses are presented in Table 2. All

TABLE 2

ANALYSIS OF VARIANCE OF DIFFERENCES BETWEEN
TEST PERFORMANCE UNDER STANDARD AND ROLE
PLAYING INSTRUCTIONS

Source of Variance	<i>df</i>	Rorschach		MMPI	
		<i>MS</i>	<i>F</i>	<i>MS</i>	<i>F</i>
Chronicity of Illness	1	2.04	4.57*	349.09	1.07
Rapidity of Improvement	1	0.51	1.14	592.24	1.82
Interaction	1	0.26	0.59	180.08	0.55
Within Groups	50	22.33		16,253.83	

* *p* < .05.

of the results are in the predicted direction, but only one achieved statistical significance. The acute group improved its Rorschach performance significantly more in playing the normal role than did the chronic group.

DISCUSSION

Studies of changes in Rorschach protocols between two test administrations in the absence of instructions to play a specified role suggest that signs of psychopathology visible on the first protocol are equally clear on the second. Griffith (1951) tested a sample of four patients with a diagnosis of Korsakoff syndrome whose memory was sufficiently impaired so that they did not recall the first test situation when repeating the Rorschach 1 day afterwards. In his summary of these four cases Griffith comments that autistic original percepts reliably characterized the individual and that this kind of responsivity is consistent on both protocols. Holzberg and Wexler (1950) studied a group of schizophrenics who were tested twice with the Rorschach test with an interval of 3 weeks between adminis-

trations. Significant statistical changes could not be demonstrated for any Rorschach factor from test to retest. They suggested that chronic schizophrenics are highly consistent. These studies indicate that one can reasonably expect reliable performance from test to retest even within a psychotic population in the absence of special role taking instructions.

The description of the schizophrenic as a person lacking a concept of a "generalized other," offered by Sarbin (1943) and elaborated by Helfand (1956), is consistent with the performance of the chronic schizophrenics of the present study. This description is less appropriate, however, for the acute schizophrenics since the Rorschach results suggest that this group has some conception of the normal role and can differentiate it from schizophrenia. This is particularly true of those acute schizophrenics who subsequently showed rapid clinical improvement.

Comparison of the results on the Rorschach and MMPI indicates that the Ss of the study were able to reduce their schizophrenia scores on the former but not on the latter. It might be assumed that a well-structured task such as the MMPI Sc scale would prove more responsive to role playing than a less structured one such as the Rorschach. Scores obtained by the experimental group were in the range typically found by other investigators when studying schizophrenics, which suggests a certain degree of validity in the present findings insofar as the standard instruction MMPI is concerned. Sorting the 78 cards proved to be a long and laborious task, however, and it was not easy to maintain all Ss' interest in the test or to ensure a proper set. Furthermore, a number of items were found to be worded in a complex and possibly confusing fashion. (Example: "I do not often notice my ears ringing or buzzing.") Answering such items under standard instructions seemed difficult for a number of Ss; with the added operation required under role playing instructions, the difficulty seemed to be compounded. Whether the present results reflect these difficulties, which might have led to random sorting on the role playing task, or whether the results suggest a true inability of schizophrenics to predict how normals would respond to these

test items, remains an unanswered question here.

Because results with the MMPI in the present study are at variance with the results reported by Grayson and Olinger (1957), a sample of 15 schizophrenic women, none of whom had been included in the study reported here, were administered the entire MMPI under both standard and role taking instructions. An analysis of this small sample confirms the results found in the present study. No significant changes were found in the apparent ability of this small sample to reduce signs of psychopathology on the MMPI under role taking instructions. In a sample of 14 schizophrenic men tested with the MMPI under standard and role taking instructions, there was again no significant reduction in signs of psychopathology from the first to the second test administration on any MMPI variable. In comparing the sample of cases used in the present study with the sample used by Grayson and Olinger, it seems possible that their sample consisted of cases who did not have as extensive psychopathology as the sample used in the present investigation. Furthermore, it is possible that there is a difference in general language facility. These two differences may account for the different findings. The ability to take the normal role appears to be a variable trait which is partially related to severity and longevity of psychopathology and to subsequent clinical history. Level of intelligence, however, especially language facility, may also be related to this trait.

SUMMARY

In the present study an attempt was made to throw further light on role playing in schizophrenia. On the basis of previous research, it was hypothesized that acutely ill schizophrenics would be better able to play the normal role than chronically ill ones, and that whether acutely or chronically ill, schizophrenics who subsequently improved would be better able to play the normal role than those who did not. Although the results were all in the predicted direction, they did not generally achieve high statistical significance.

REFERENCES

- CROW, W. J. The accuracy of "experts" in making judgments about others. Paper read at Western Psychological Association, San Diego, April 1959.
- DYMOND, ROSALIND F. A preliminary investigation of the relation of insight and empathy. *J. consult. Psychol.*, 1948, 12, 228-233.
- DYMOND, ROSALIND F. A scale for the measurement of empathic ability. *J. consult. Psychol.*, 1949, 13, 127-153.
- DYMOND, ROSALIND F. Personality and empathy. *J. consult. Psychol.*, 1950, 14, 343-350.
- GRAYSON, H. M., & OLINGER, L. B. Simulation of "normalcy" by psychiatric patients on the MMPI. *J. consult. Psychol.*, 1957, 21, 73-77.
- GRIFFITH, R. M. Test-retest similarities of the Rorschachs of patients without retention, Korsakoff. *J. proj. Tech.*, 1951, 15, 516-525.
- HASTORF, A. H., & BENDER, I. E. A caution respecting the measurement of empathic ability. *J. abnorm. soc. Psychol.*, 1952, 47, 574-576.
- HELFAND, I. Role taking in schizophrenia. *J. consult. Psychol.*, 1956, 20, 37-41.
- HOLZBERG, J. D., & WEXLER, M. The predictability of schizophrenic performance on the Rorschach test. *J. consult. Psychol.*, 1950, 14, 395-399.
- JACKSON, W., & CARR, A. C. Empathic ability in normals and schizophrenics. *J. abnorm. soc. Psychol.*, 1955, 51, 79-87.
- LINDGREN, H. C., & ROBINSON, JACQUELINE. An evaluation of Dymond's test of insight and empathy. *J. consult. Psychol.*, 1953, 17, 172-176.
- McCLELLAND, W. A. A preliminary test of role playing ability. *J. consult. Psychol.*, 1951, 15, 102-108.
- NORMAN, R. D., & AINSWORTH, PATRICIA. The relationship among projection, empathy, reality, and adjustment, operationally defined. *J. consult. Psychol.*, 1954, 18, 53-58.
- SARBIN, T. R. The concept of role taking. *Sociometry*, 1943, 6, 273-285.
- SCHAFER, R. *The clinical application of psychological tests*. New York: International Univer. Press, 1948.

(Received December 22, 1959)

THERAPIST-PATIENT RELATIONSHIPS AND OUTCOME OF PSYCHOTHERAPY

MORRIS B. PARLOFF¹

National Institute of Mental Health

An assumption underlying most forms of psychotherapy is that the relationship between the therapist and his patient is the vehicle for therapeutic change. More specifically, the benefits from therapy are believed to vary directly with the quality of the therapist-patient relationship (Betz & Whitehorn, 1956; Freud, 1949; Rogers, 1951; Snyder, 1959). It is frequently assumed that the fact of a patient's remaining in treatment may be interpreted as evidence of the "goodness" of the relationship and therefore of the probability of an ultimately successful outcome.

These widely held beliefs have not, however, gone unchallenged. Eaton (1959) recently warned that a "good relationship" may indeed interfere with therapeutic outcome. He stated that a "good relationship may help influence the client to become dependent on such help and to continue seeking it," thereby defeating the therapeutic goal of helping the client to achieve autonomy. Redl and Wineman (1951) also pointed out the potential limitations of a seemingly good relationship. They stressed that the therapist who establishes a close, warm, and permissive relationship with a patient may find himself occupying the non-therapeutic role of "friend without influence." These contradictory viewpoints may be due to the fact that the investigators used differing definitions of the concept good relationship. What is required is a more explicit definition of the therapist-patient relationship concept and the systematic testing of it. To

date there have been very few experimental tests made regarding the association between favorable outcome and quality of the therapist-patient relationship (Heine, 1950; Holt & Luborsky, 1952; Snyder, 1953).

There are many theoretical frames of reference from which the concept of relationship may be viewed, yet, according to Fiedler (1950), these differences may readily be subsumed under one general description of the "ideal therapeutic relationship." He reported that therapists of diverse theoretical persuasions revealed a remarkable degree of agreement in characterizing the ideal therapeutic relationship. The very concept of the ideal therapeutic relationship appears however to violate the clinician's belief that to be effective a relationship must be adapted and modified to meet the particular needs of a given patient. An examination of the Fiedler instrument is reassuring since the descriptive statements are written at such a high level of abstraction as to encompass the relationship needs of most patients as well as most non-patients. This study, therefore, employs Fiedler's view of the therapeutic relationship to provide an operational definition of this elusive concept.

This report describes an effort to test, in a group therapy setting, the correlations between patient-change, remaining in treatment, and quality of therapeutic relationship. Since the definition of the construct "therapeutic relationship" has not been widely agreed upon, and no criteria have been accepted as valid, this study is to be viewed as a test of the concept's construct validity (Cronbach & Meehl, 1955). It is further recognized that the construct validity of the outcome criteria employed here also may be regarded as under study.

¹ This study was conducted at the Henry Phipps Psychiatric Clinic of the Johns Hopkins Hospital, Baltimore, Maryland. The author expresses grateful acknowledgment to J. D. Frank, E. Ascher, H. Kelman, D. Rosenthal, E. Nash, and A. R. Stone for their cooperation and many helpful suggestions.

The concept of the "therapeutic process" presupposes that psychotherapy proceeds in a discernibly systematic step-wise fashion. Therefore, some investigators believe that certain kinds of intermediate change may be viewed as harbingers of significant benefit to the patient. The investigator who accepts this idea ascribes to a variety of phenomena the status of "enabling" or intermediate conditions necessary for beneficial change. Unfortunately, the relationship between the alleged roadmarkers and the destination is not as yet established. In group therapy, for example, it may be encouraging to the therapist to note increased group cohesiveness, evidences of group support and stimulation, establishment of multiple transferences, recall of repressed material, resolution of transferences, etc. That such phenomena do not invariably eventuate in clinical improvement will be conceded even by the most ardent group therapist. The author decided, therefore, to concentrate on criteria that related to ultimate goals, such as providing symptomatic relief and improving social functioning, rather than intermediate goals. The assumption that "improvement" is a unitary phenomenon is questionable (Kelman & Parloff, 1957). This is especially the case where improvement is "less than complete recovery." This broad category unfortunately includes a considerable proportion of all patients treated. If, then, improvement cannot be discussed in global terms, it is necessary to specify the various criteria and measures.

The three criteria of improvement adopted in this study are based on the work of Kogan and Hunt (1950) and Miller (1951). These criteria are: Comfort, Effectiveness, and Objectivity. The first two criteria, Comfort and Effectiveness, are based on the belief that the general aim of psychotherapy is to ameliorate the patient's suffering and to restore him to an effective level of functioning in the community. Comfort was defined in terms of symptoms or feelings which had caused distress. Effectiveness was defined as the degree of competence with which the patient managed to fulfill his own needs and desires as well as those of others. With the third criterion, Objectivity, an attempt was made to take cognizance of a value which is shared by

a number of quite different psychotherapeutic approaches. All assert that the better the individual understands himself the freer he will be to react appropriately to conditions arising in his current life. Objectivity is not an end-point per se but a generally accepted means to an end.

That the patient remain in treatment is a necessary but not sufficient condition for psychotherapy to be effective. The amount of time necessary for change to occur varies from patient to patient. The patient may remain in treatment and yet fail to improve. Although a patient who drops out of therapy may have derived considerable benefit, his departure may preclude any objective assessment of this benefit. The factors that determine whether a patient will remain in treatment may or may not coincide with those which determine whether he will improve if he does remain.

In the present study it was hypothesized that changes evidenced in Comfort, Effectiveness, and Objectivity are related to the quality of the therapeutic relationship. It was also hypothesized that remaining in treatment is similarly a function of the goodness of the relationship.

METHOD

The eight instruments used in this study to measure the criteria (Comfort, Effectiveness, and Objectivity) will be described only briefly. A fuller description may be found elsewhere (Kelman & Parloff, 1957).² Since the therapy goals to be reported were the amelioration of discomfort and the modification of ineffectual behavior, the scales measuring Comfort and Effectiveness were reversed to measure instead the degree of "Discomfort" and "Ineffectiveness."

Judgments regarding the patient's Discomfort, Ineffectiveness, and Objectivity were made by research teams composed of a psychiatrist, social worker, and psychologist. Specially designed "evaluation" scales were filled out independently by each member of the research team in describing each patient. These judges then met to discuss their ratings and to arrive at an overall single staff rating for the scale measur-

² Copies of measures used in the evaluation of psychotherapy have been deposited with the American Documentation Institute. Order Document No. 6464 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance \$1.75 for microfilm or \$2.50 for photocopies. Make checks payable to: Chief, Photoduplication Service, Library of Congress.

ing each criterion. In addition to these staff ratings, two measures of Discomfort were obtained from the patients, one measure of Ineffectiveness was derived from ratings made by group members of each other, and two measures of Objectivity were obtained by (a) comparing the patient's self-description with an independent staff observer's description of him, and (b) determining the accuracy of the patient's predictions of the ratings he would receive from each of his fellow patients. The staff members who participated in completing the staff evaluation scales did not act as judges of the therapeutic relationship for the same patients except in two cases.

Measures of Discomfort

1. *Self-Satisfaction Q sort (Patient)*: This attempted to tap the degree of congruence between the patient's perceived self and ideal self regarding behavior in the group therapy situation. A 60-item "perception" Q sample was employed. It is based on Bion's (1950) group interaction concepts.

2. *Symptom Disability Checklist (Patient)*: This is a modification of the Cornell index. Forty-one items referring to psychic or somatic complaints were rated by the patient in terms of the relative distress they caused him during the week preceding testing.

3. *Discomfort Evaluation Scale (Staff)*: This consists of items describing 10 areas of interpersonal discomfort. The scale was filled out independently for each patient by three staff raters: psychiatrist, social worker, and psychologist. These judges then met to discuss the ratings and to agree on an overall single staff rating on the basis of their combined clinical judgment.

Measures of Ineffectiveness

1. *Ineffectiveness Evaluation Scale (Fellow Patients)*: Each patient rated each of the other patients in his therapy group on three dimensions: the extent to which the rater respected another patient's ideas and opinions, regarded him as a group leader, and desired to be friends with him. Ratings on each dimension were made on a four-point scale and were reported individually as measures of Respect, Leadership, and Friendship. The average overall rating a patient received on the three measures was also computed.

2. *Ineffectiveness Evaluation Scale (Staff)*: This scale consists of 15 items in which the patient's creativity, productivity, and fulfillment of social roles are rated. The ratings concern the degree of appropriateness of the behavior and the frequency with which it occurred in relation to significant persons in the patient's home and community life. The above mentioned staff members independently completed this form. On the basis of a conference a unified staff rating was made.

Measures of Objectivity

1. *Objectivity Q sort (Patient-Observer)*: Objectivity was measured by the degree of congruence be-

tween the patient's description of his group behavior and the staff observer's description of the patient's behavior in the group. The Q sort items were the same as those used for measuring self-satisfaction.

2. *Objectivity Evaluation Scale (Patient-Fellow Patient)*: In completing the group questionnaire previously described, patients were asked to predict the ratings which they would receive from each of their fellow group members. The average discrepancy between the ratings each patient expected from each fellow patient and the ratings he actually received was computed for each of three areas: Respect, Friendship, and Leadership.

3. *Objectivity Evaluation Scale (Staff)*: This scale consists of four items attempting to measure the accuracy of the patient's perceptions of his own behavior and the behavior of others. Staff ratings were made independently and then combined into a single staff rating by the conference method.

Except for the Symptom Disability Checklist, which was completed prior to therapy, all initial testings were made immediately after the fourth group session. All measures were repeated following the twentieth group meeting.

Drop-Outs

Any patient who left the group prior to the twentieth session without his therapist's approval was considered to have terminated prematurely. Four of the 21 patients were so designated. By the end of the experimental period each patient had attended an average of 9.6 sessions. The attendance ranged from 5 to 12 sessions.³

Therapeutic Relationship

The technique developed by Fiedler (1950) was employed. He had the 75 statements in his Relationship Q Sample sorted by members of various schools of therapy. On this basis, an ideal therapeutic relationship standard was developed. Twenty-five items concerned the therapist's ability to communicate with and to understand the patient, 25 described the "emotional distance" between the therapist and patient, and the remaining 25 dealt with questions of "status" as reflected in the therapist's behavior toward the

³ Prior to assignment to one of three therapy groups, each of the 21 patients had been "screened" by exposure to a 6-week orientation group. Experience with group therapy had indicated that more than one-third of the patients dropped out of therapy by the end of the fifth session. This involved a loss of the time invested in initial evaluations of such patients. The aim of the orientation group was to expose patients to a group experience similar to that which they might experience in the actual therapy situation. It was hoped that patients who survived six sessions might then tend to remain in group therapy. The selection process did, in fact, act to increase the proportion of patients remaining beyond the fifth hour in therapy. Only one dropped out of treatment by the fifth hour.

patient. In the present study, the 75 items were used by the observers to describe the relationship between therapist and patient.⁴ These arrays were then correlated against the Fiedler ideal therapeutic relationship standard. The higher the correlation with the standard, the "better" the relationship. Three trained observers were used as judges to describe the therapeutic relationship established by two therapists. After a preliminary practice period, the interjudge reliability in describing the same therapist-patient interaction was found to be substantial. The correlation between the relationships described by pairs of judges in observing 19 patient-therapist interactions was .92.⁵ One of the three judges was assigned to each of the three groups and attended all group meetings during the experimental period of 20 weeks. Each judge described the relationships established by the therapist with each patient in the group after the second meeting, the twelfth meeting, and the twentieth meeting. Each of the three descriptions was correlated against the standard and the overall relationship was described as the mean of the three correlations.⁶

The sample consisted of 21 psychoneurotic patients, 10 male and 11 female. Fourteen of the patients were classified as "psychoneurotic disorders," five as "personality disorders," one as "psychotic disorder," and one as "transient situational personality disorder." They were randomly assigned to three groups, ranging in size from six to eight. A treated 13 patients, 6 in one group and 7 in the other. B

⁴ Fiedler's concept of the ideal therapeutic relationship was derived primarily from experiences in individual therapy. It was necessary, therefore, to determine whether the group therapists in the present study conceived of the ideal therapeutic relationship in a similar fashion. Each therapist was asked to describe, by means of the 75-item Q sort, his conception of the ideal therapeutic relationship. These arrays were correlated with the Fiedler standard. It was found that A's and B's ideals correlated .86 and .88, respectively, with the Fiedler "ideal." It was concluded, therefore, that the aims of these group therapists were sufficiently consonant with the Fiedler standard to permit its use as the criterion for measuring the goodness of relationships.

⁵ Since the data collection involved the use of independent group observers, the question of the interrater and intrarater reliability is an important one. Although the reliability measures described here appear to be adequate, data are available only for the initial period of the study. Since no further attempt to check on the reliability of the judges was made during the period of the study, there is no direct evidence that judges continued to describe the therapeutic relationship in a consistent manner throughout the experiment. Indirect evidence on this point is found in the fact that the patients' descriptions of their relationships with their therapists correlated substantially with those ascribed to them by the judges ($\rho = .79$).

⁶ The necessary z transformations were made.

treated one group of 8 patients. The data reported are based on the first 20 sessions of each group and are, therefore, limited to the early period of treatment. Of the initial 21 patients, 14 completed all experimental procedures by the close of the 20-week period.⁷

Each group met for an hour and a half once a week. The form of therapy was largely interpretive with the focus on the immediate interaction of the patients with each other and with the therapist.

The two therapists qualified as "experts" as described by Fiedler, i.e., each had completed prescribed training, had been a practicing therapist for a minimum of 5 years, and was considered an expert by other therapists within his school.

RESULTS

To determine whether improvement varies with the quality of the therapeutic relationship, product-moment correlations were computed between the Fiedler ideal therapeutic relationship scores and each of the 14 change scores.⁸ Inspection of the therapeutic relationship scores revealed that the four patients treated by B had each achieved relationships which were higher than any established by

⁷ In addition to the four drop-outs already mentioned, one patient left treatment when her husband's job was transferred out of the city, and one failed to complete all evaluation procedures. Another patient was excluded from the study when it was learned that he had supplemented group therapy with intensive individual therapy. The effectiveness of the group therapy relationship was, therefore, confounded with the individual therapy relationship.

⁸ To determine whether the evaluation measures were initially related to the quality of therapeutic relationships subsequently established, correlations were computed between initial scores on each of the 14 measures and the overall mean therapeutic relationship scores. The correlations obtained did not differ significantly from zero. The range was from .126 to -.452. (In order for a correlation with 12 degrees of freedom to be significant at the .05 level of confidence, a correlation of .532 is required.)

To further test whether the therapeutic relationship was associated with initial scores on these evaluation measures, the initial scores of the seven patients who had therapeutic relationships above the group median were compared with the seven patients whose therapeutic relationship scores fell below the median. None of the group differences as tested by the t test attained statistical significance. In view of the apparent lack of association between the 14 initial evaluation scores and the quality of the subsequent therapeutic relationships established, we were justified in computing correlations between therapeutic relationships and the difference scores between the initial and final evaluation scores.

the 10 patients treated by A. In effect, the therapeutic relationships established by each therapist with his patients appeared to come from different "populations" of relationships. This suggested that the influence of the therapeutic relationship variable, which is at issue in this investigation, may be confounded with other variables related to the individual therapist.

The correlations between therapeutic relationship and the measures of change were therefore computed independently for A's patients and for B's patients. To obtain the best estimate of the true mean correlations for the total sample a pooled correlation (\bar{r}) was computed.⁹ As may be seen in Table 1, 3 of the 14 mean correlations differed significantly

⁹ To determine whether the patients of each therapist differed on their initial evaluation measure scores, a Mann-Whitney *U* test was computed for each of the 14 measures. No significant differences were found.

TABLE 1
POOLED MEAN CORRELATIONS BETWEEN CHANGE AND
MEAN THERAPEUTIC RELATIONSHIP
(Computed for 10 Patients Treated
by A and 4 by B)

Measure	Pooled \bar{r} (<i>N</i> = 14)
Discomfort	
A. Self-Satisfaction <i>O</i> sort (Patient)	.037
B. Symptom Disability Checklist (Patient)	.669**
C. Discomfort Evaluation Scale (Staff)	.012
Ineffectiveness	
A. Ineffectiveness Evaluation Scale (Fellow Patients)	.103
1. Respect	.613*
2. Leader	.217
3. Friend	.460
4. Overall Total	
B. Ineffectiveness Evaluation Scale (Staff)	.133
Objectivity	
A. Objectivity <i>O</i> sort (Patient- Observer)	.521
B. Objectivity Evaluation Scale (Patient-Fellow Patient)	
1. Respect	.183
2. Leader	-.161
3. Friend	-.013
4. Overall Total	.082
C. Objectivity Evaluation Scale (Staff)	.669**

Note.—The direction of "negative" scales has been reversed so that a positive correlation between a criterion and relationship indicates that the greater the relationship the greater the improvement; a negative correlation indicates a negative relationship between relationship and improvement.

* r significant at the .05 level, one-tailed test.

** r significant at the .01 level, one-tailed test.

TABLE 2
PATIENT-THERAPIST RELATIONSHIPS (\bar{z}) ESTABLISHED
WITH A AND B

	A		B
	Group I	Group II	Group III
	.51 ^a	.74	1.28
	.23	.71	1.10
	.15	.63	1.06
	.10	.62	1.00 ^b
	.09	.53	.99
	.04	.24 ^c	.81 ^c
		.17 ^d	.73 ^c
			.63 ^c
Mean	.19	.52	.95
<i>SD</i>	.17	.225	.214
<i>N</i>	6	7	8

^a Moved out of city.

^b Failed to complete evaluation procedures.

^c Terminated prematurely.

^d Supplemented group therapy with simultaneous individual therapy.

from zero.¹⁰ The findings indicate that the more closely the therapeutic relationship approximated the ideal relationship the greater the increase in the patient's Objectivity (as evaluated by the staff, $\bar{r} = .67$, $p < .01$); the greater the increase in group Effectiveness—Leadership (as derived from ratings by fellow patients, $\bar{r} = .61$, $p < .05$); and the greater the relief from symptomatic Discomfort (as reported by the patient, $\bar{r} = .67$, $p < .01$). It is noted that the correlation between therapeutic relationship scores and Objectivity as measured by the *O* sort falls just short of reaching an acceptable level of significance ($\bar{r} = .52$, while an r of .53 is required for $p < .05$).

These findings indicate that the quality of the therapeutic relationship does vary on certain measures with patient-change in the areas of Objectivity, Effectiveness, and Comfort when the therapist variable is controlled.

That an association exists between the quality of the therapeutic relationship and the incidence of drop-outs is strongly suggested (see Table 2). When the eight patients initially assigned to B's group were ranked according to the quality of the therapeutic relationship established, it was found

¹⁰ Since the direction of the correlation was predicted, a one-tailed test was applied.

TABLE 3
MEAN DIFFERENCE IN OVERALL
GROUP RELATIONSHIPS

Comparison	Mean <i>z</i> Differ- ence	<i>df</i>	<i>t</i>	<i>p</i>
Group I vs. II	.33	11	2.94	.02
Group I vs. III	.76	12	7.17	.001
Group II vs. III	.43	13	3.80	.01

that those patients who had remained in treatment occupied ranks 1 through 5. Those who terminated prematurely were in rank order positions 6 through 8. In investigating the rank order position of the single individual who dropped out of one of A's groups, it was found that of a group of seven the terminator had occupied position Number 6. Moreover, the patient in rank order position Number 7 was found to have supplemented his group therapy experience with intensive individual psychotherapy without having notified the group therapist. Thus, the five who dropped out or found it necessary to supplement group treatment appear to have had the poorer relationships when compared to the other members of the particular group to which they were assigned. An inspection of the group in which no member dropped out showed that the relationships were quite uniform and, incidently, very low. The variance within this group tended to be smaller than in the other groups. The mean relationship in this group (Group I) was the lowest of the three groups (see Table 3). Thus the group having the poorest relationships remained intact while the group having the highest relationships lost 38% of its members. From Table 2 it appears that the quality of the therapeutic relationship is related to premature termination; however, the absolute size of the relationship score appears to be less important than the terminator's relationship score relative to those of other members of his group.

In designing the study it was decided to use observers to describe the relationships since it was assumed that patients' perceptions of their relationships with the therapist would

be subject to the distorting influence of transference. However, the fact that the patients' premature termination of therapy appears to be related to the quality of relationships, as evaluated by observers, implies that the patients experienced their relationships much as described by these judges. To investigate this apparent concordance, an attempt was made to determine the degree of agreement between the observer's and the patient's descriptions of the relationships. At the end of the experimental period the 14 remaining patients were asked to describe their relationship with the therapist by using the Fiedler Q sort. Various items were modified to clarify technical terms. Patients were assured that the data they furnished was confidential and would not be relayed to their therapists. Their descriptions were then correlated with the ideal to derive "relationship scores." Each patient's relationship score was then correlated with his mean relationship score as provided by the observer. The rho correlation was found to be .79 ($p < .01$). This finding strongly suggests that in these groups patients were quite objective in perceiving their relationships with their therapists. Transference distortions as measured here appear to have played a surprisingly small role.

Although the focus of the paper has been on the effect of the patient-therapist relationship on the outcome of group therapy, it is recognized that the patient-patient relationships also play a significant role. In the sample of patients studied it was found that those who established better relationships with their therapist reported a significantly greater inclination to perceive other group members as being socially attractive (Ineffectiveness Evaluation Scale-Friendship) than those who formed poorer relationships with their therapist.

DISCUSSION

The findings of this study appear to provide limited support for the hypothesis that patient-change in psychotherapy is related to the quality of the therapeutic relationship established. Of a total of 14 change measures, 3 revealed significant correlations with therapeutic relationship scores. It is of interest that one measure of each criterion of im-

Improvement—Discomfort, Ineffectiveness, and Objectivity—showed this association. The data indicate that the better the patient-psychotherapist relationship, the greater the symptomatic relief experienced by the patient, the more likely it was that fellow group members would describe the patient as having become more dominant (leader), and the greater the increase in Objectivity attributed to the patient by the research staff. Since the treatment period included in this study was arbitrarily limited to the initial 20 weeks, it is not clear whether the associations described here characterize only the early stages of psychotherapy. It is not known whether these correlations are maintained, diminished, or increased, or whether additional correlations will be found between other outcome measures and the therapeutic relationship scores in later periods of psychotherapy. It is possible that not all the behaviors measured in this study are modifiable at the same rate. The probability that a significant correlation would be found with the therapeutic relationship scores was not necessarily equal for each of the 14 measures. However, the researcher did assume that each measure had face validity and that therefore the hypothesis could reasonably be tested against each of these 14 measures.¹¹

¹¹ Experience with the Objectivity Evaluation Scale (Patient-Fellow Patients) indicates that it required the raters of the four measures contained in this scale to make rather complex judgments. The ostensible task for the patient was to predict the rating he would receive from each patient on each of three measures: Respect, Leadership, and Friendship. In order to evidence Objectivity he had to be able to predict not only how he was perceived by each of his fellow patients but also the rating which each would be willing to assign to him. The rater is frequently keenly aware that the scores which he assigns to others will also provide the investigator with information about the rater. Therefore, the ratings are frequently intended primarily as a communication to the examiner rather than as an objective report of the raters' experiences and feelings about fellow patients. Some patients wish to be seen as warm and friendly and therefore assign high scores fairly indiscriminately to their fellow patients. Others wish to be seen as aloof and independent of the group members and therefore assign low scores. In addition, some patients appear reluctant to reveal warm feelings concerning group members of the opposite sex and therefore tend to minimize these ratings.

Statistically significant changes for the group as a whole, over the 20-week period, occurred only on three measures: Symptom Disability Checklist, Discomfort Evaluation Scale (Staff), and Ineffectiveness Evaluation Scale (Staff). Only one of these change measures was found to be significantly correlated with therapeutic relationship scores. Although it is possible that individual patients experienced real changes even if the overall sample did not, it is equally possible that the change scores used in this study may represent measurement error, particularly in those instances where a lack of correlation with therapeutic relationship scores was shown. In those instances where no significant change occurred, the possibility of measurement error must be considered a possible explanation for the lack of correlation between change scores and therapeutic relationship scores.

The question arises as to the explanation of the finding that one measure of each criterion showed a significant association with the quality of the therapeutic relationship, but other measures of the same criterion failed to show similar associations. One condition under which such findings could have occurred is that each of the three hypothesis-supporting measures was independent of the other measures of a given criterion. To the degree that the various measures are independent of one another they may be expected to vary independently with the quality of the therapeutic relationship. A second condition which might account for these findings is that the measures were related to each other in a complex fashion—i.e., two measures may be correlated under some circumstances, and not correlated, or even negatively correlated, under other circumstances. To determine the degree and nature of the association between the three criterion-supporting measures and the other measures of the relevant criterion, the initial scores on these measures were intercorrelated. Thus scores on the Symptom Disability Checklist were correlated with the other two Discomfort measures, Objectivity Evaluation Scale (Staff) scores were correlated with the other five Objectivity measures, and Leader scores on the Ineffectiveness Evaluation Scale were correlated with the other four Ineffectiveness measures. The con-

dition of independence of measures was found to be the case with the Objectivity group of measures. No significant correlations were found between the initial scores of the Objectivity Evaluation Scale (Staff) and the remaining Objectivity measures. Since the measures apparently tap different aspects of the criterion, the fact that change on one measure correlates with relationship scores does not lead to the expectation that change on other measures will also be associated with relationship scores.

The possibility that some contamination of measures had occurred on the Objectivity Evaluation Scale (Staff) must be considered since in two instances one member of the three-man team of staff raters had also been the observer who described the therapeutic relationships. This fact need not cast serious doubt on the authenticity of the correlation between changes on Objectivity Evaluation Scale (Staff) and quality of therapeutic relationship for the following three reasons: (a) Staff ratings were determined by conferences of three staff members. There is no evidence that the opinion of any one team member was weighted disproportionately. (b) Even under the most unfavorable circumstances only 2 of the 14 cases could have been affected by the possible contamination of measures. (c) If the staff ratings and the judgments regarding therapeutic relationships had been spuriously correlated due to contamination of measures, the same association would also be anticipated with the other two staff ratings on Ineffectiveness and Discomfort. No such evidence is found.

The condition of "complexity of association" between measures appears to apply to the cases of Ineffectiveness and Objectivity measures.

In the case of Leadership, Ineffectiveness Evaluation Scale (Fellow Patients), it was found that it failed to correlate significantly with any other measure of Ineffectiveness with the exception of the Overall Total score. This is to be expected since the Overall Total measure is simply the sum of Leadership, Respect, and Friendship scores. Leadership scores, however, were not found to correlate with either Respect or Friendship. Apparently a patient's dominant group behavior

did not win him the respect or friendship of his peers. Since the Overall Total Evaluation Scale score is made up of components which are independent of each other, it is not surprising that changes on this measure do not correlate with therapeutic relationship scores, despite the fact that Leadership change scores do show a significant association with therapeutic relationship scores.

The Symptom Disability Checklist, a measure of Discomfort, was found to correlate significantly (.55) with Self-Satisfaction *Q* sort. The expectation that changes in Self-Satisfaction *Q* sort scores might, like the Symptom Disability Checklist change scores, correlate with therapeutic relationship scores was not supported. An analysis of the Self-Satisfaction *Q* sort data revealed that two quite opposite changes had occurred. Those patients who initially reported very high self-satisfaction appeared to become less content with themselves, while those who initially showed the greatest discontent tended to show greater self-satisfaction as therapy progressed. This is not a simple regression toward the mean for it was found that the patients who initially indicated that their group behavior very closely approximated their ideal behavior were the ones who showed rather poor Objectivity. As these individuals became increasingly aware of their actual behavior, this was reflected in their description of their group behavior as being less in accord with their rather stable ideals. As a result of this shift the initial correlation between self-satisfaction and symptomatic comfort was dissipated.

The fact that a positive relationship was found between the quality of therapeutic relationship and three measures does not, of course, permit one to assign direction to this association. In the case of Leadership Ineffectiveness Evaluation Scale, for example, two equally plausible interpretations can be made: patients who established the better relationships with their therapists were able to become more assertive and dominant in their therapy groups; or, the therapist tended to relate to those patients who manifested increasing evidences of leadership and dominance in the group. Despite the fact that the findings are consistent with the general hypothesis that improvement follows upon the

establishment of a good relationship, other alternative interpretations must be considered. The possibility that the relationship established may be associated with characteristics of the patient is not ruled out by the absence of a significant correlation between initial scores on the 14 measures used in this study and the quality of the subsequent therapeutic relationship. On the contrary, there is compelling evidence that the therapist's perception of his patients is intimately associated with the quality of the relationship he is able to establish with them (Parloff, 1956). Such correlation was evidenced not only in the initial 4 weeks of therapy, but throughout the 20-week experimental period (Parloff, 1953).

It may be possible, for example, that one of the characteristics of the expert group therapist is his ability to identify the therapeutic potential of his patients. As a consequence, he may then direct his attention toward effecting a positive relationship with those patients with whom he feels he can be most useful. Indeed, he may recognize and attempt to relate to those individuals who tend to improve seemingly independent of specific therapeutic efforts.

Although the findings support the notion that Fiedler's instrument has a measure of construct validity, the *Q* sort may offer only a partial or even a superficial definition of the therapeutic relationship. The ideal therapeutic relationship standard defined by Fiedler seems to describe conditions essential to any meaningful social relationship, independent of therapeutic intent. This type of relationship may be therapeutic per se; it also is possible that this relationship may be principally a prerequisite condition for the establishment of an as yet undefined and unmeasured therapeutic relationship. Such a relationship may involve the utilization of specialized techniques and procedures which the therapist may regard as essential for treatment, e.g., analysis of transference, free association, dream interpretation, reliving of earlier emotional experiences, etc. One interpretation of the finding that patients who establish the better relationships with their therapists tend to perceive others as being more socially desirable may be that patients in a group take

their cue in relating to each other from the quality of the therapist's relationships with them. A more parsimonious explanation is that both therapist and patients react similarly to a given situation.

SUMMARY

This study reports an attempt to determine whether an association exists between the construct "therapeutic relationship" and outcome of treatment in a group therapy setting. The quality of the therapeutic relationship was measured by Fiedler's ideal therapeutic relationship *Q* sort. Three criteria of improvement were used: Comfort, Effectiveness, and Objectivity. These criteria were measured by 14 scales. In addition, a study was made of the therapist-patient relationships established with patients who terminated therapy prematurely.

The sample included 21 patients, 13 of whom were treated by one group therapist and 8 by another. The experimental treatment period was limited to 20 weeks, at which time outcome data were available on 14 patients.

Patients who established better relationships with their therapist tended to show greater improvement than those whose relationships with the same therapist were not as good. In computing the overall pooled mean correlations between the therapeutic relationship scores and change measures for 14 patients, significant correlations were found on three measures: increased Objectivity (staff evaluation), increased Effectiveness (group leadership), and increased self-ratings of Comfort (symptomatic relief).

Premature termination of therapy by a patient appears to be related to his perception of the "goodness" of the relationship he has established with his therapist *relative* to the general level of patient-therapist relationships within his group. Individuals having the poorer relationships in a group tended to drop out of therapy irrespective of the *absolute* goodness of their therapeutic relationship.

The hypotheses postulating that benefit from psychotherapy and incidence of premature termination were associated with the goodness of the individual patient-therapist

relationship tended to be supported. These findings give limited support to the validity of the concept "therapeutic relationship" as defined by Fiedler.

REFERENCES

- BETZ, BARBARA J., & WHITEHORN, J. C. The relationship of the therapist to the outcome of therapy in schizophrenia. *Psychiatric res. Rep.*, 1956, No. 5.
- BION, W. R. Experiences in groups. V. *Hum. Relat.*, 1950, 3, 3-14.
- CRONBACH, L. J., & MEEHL, P. E. Construct validity in psychological tests. *Psychol. Bull.*, 1955, 52, 281-302.
- EATON, J. W. The client-practitioner relationship as a variable in the evaluation of treatment outcome. *Psychiatry*, 1959, 22, 189-195.
- FIEDLER, F. E. The concept of an ideal therapeutic relationship. *J. consult. Psychol.*, 1950, 14, 239-245.
- FREUD, S. *An outline of psychoanalysis*. New York: Norton, 1949.
- HEINE, R. W. An investigation of the relationship between changes and responsible factors as seen by clients following treatment by psychotherapists of the psychoanalytic, Adlerian and non-directive schools. Unpublished doctoral dissertation, University of Chicago, 1950.
- HOLT, R. R., & LUBORSKY, L. Research in the selection of psychiatrists: A second interim report. *Bull. Menninger Clin.*, 1952, 16, 125.
- KELMAN, H. C., & PARLOFF, M. B. Interrelations among three criteria of improvement in group therapy: Comfort, effectiveness, and self-awareness. *J. abnorm. soc. Psychol.*, 1957, 54, 281-288.
- KOGAN, L. S., & HUNT, J. McV. After comment. *Psychol. Serv. Cent. J.*, 1950, 2, 132-138.
- MILLER, J. G. Objective methods of evaluating process and outcome in psychotherapy. *Amer. J. Psychiat.*, 1951, 108, 258-263.
- PARLOFF, M. B. An analysis of therapeutic relationships in a group therapy setting. Unpublished doctoral dissertation, Western Reserve University, 1953.
- PARLOFF, M. B. Some factors affecting the quality of therapeutic relationships. *J. abnorm. soc. Psychol.*, 1956, 52, 5-10.
- REDL, F., & WINEMAN, D. *Children who hate*. Glencoe, Ill.: Free Press, 1951.
- ROGERS, C. R. *Client-centered therapy*. New York: Houghton Mifflin, 1951.
- SNYDER, W. U. (Ed.) *Group report of a program of research in psychotherapy*. State College, Pa.: Pennsylvania State Univer. Press, 1953.
- SNYDER, W. U. Some investigations of relationship in psychotherapy. In E. A. Rubinstein & M. B. Parloff (Eds.), *Research in psychotherapy*. Washington, D. C.: APA, 1959. Pp. 247-259.

(Received January 6, 1960)

A FACTOR ANALYSIS OF GERIATRIC ATTITUDES

WILSON H. GUERTIN¹

University of Florida

Despite current expressions of interest in geriatric patients, there is a surprising lack of attempts to evaluate the attitudes of the aged objectively. The only specific instrument available is the Activities and Attitudes Survey of Caven, Burgess, Havighurst, and Goldhamer (1949). Nor have there been factor analytic attempts to define the prominent dimensions underlying the expressed attitudes of these people. The present paper is a report of a factor analysis of such attitudes and serves as a basis for the construction of a forced-choice Geriatric Attitude Scale (Guertin & Krugman, 1961).

PROCEDURE

A total of 166 "agree-disagree" items were composed to provide a wide sampling of the attitudes of institutionalized aged. Attitudes relating to problems of adjustment were emphasized, with a few psychiatric and physical disability items included.

Forty-eight male residents of the Veterans Administration Center at Martinsburg, West Virginia, satisfactorily completed all the items. These subjects were all 60 or more years old, and none carried psychiatric diagnoses.

After excluding 16 items because responses to them were too uniformly in one direction, the remaining 150 items were subjected to a rough linkage analysis. Key sort cards provided frequencies for a four-fold contingency table for each possible pairing of items. A link was noted when the tetrachoric correlation was greater than .40 between the items. These linkages and the direction of the relationships were transferred to slips which were laid out on the floor in the form of an intercorrelation matrix. Cluster items were identified by observing that half or more of the item linkages (row entries) in a column were the same for a pair of test items.

¹ Arnold D. Krugman collaborated with the author in devising the items and supplied the data used herein. The study was conducted while the author was employed at the Veterans Administration Hospital, Knoxville, Iowa.

RESULTS

Eight important clusters were identifiable but the complexity of interrelations made it clear that only factor analysis could clarify the underlying structure. Therefore, 27 prominent cluster items were selected to provide a matrix of tetrachoric interrelations. Multiple-group factor analysis and blind rotation to oblique simple structure by the single plane method produced the factor matrix in Table 1. Items employed in the factor analysis are identified by asterisks, but the calculation of additional item loadings was necessary to provide the content for an understanding of the nature of the obtained factors.

Items in Table 1 without asterisks are those not originally factor analyzed, but for which loadings were calculated by extending the factor matrix (Cattell, 1952, p. 406). Only those items with heaviest loadings are reported here. Item descriptions are in abbreviated form and decimal points have been omitted for convenience in presentation.

It may be of interest to follow the fate of the eight clusters of variables. Three defined three of the multiple-group factors with a single variable from a fourth cluster pulled into one of the factors. Another cluster split to form two factors. The three remaining clusters failed to contribute uniquely to the factor structure.

Conventionally, factors obtained from the multiple-group factor analysis are rotated to orthogonal positions to make it possible to calculate communality and residuals. The orthogonal matrix then is rerotated blindly to oblique simple structure. Since the first matrix obtained in factoring often approximates the final oblique simple structure solution, it is of some interest to compare the two ma-

trices. The following are five items in the original factor analysis, selected from Table 1 as having the single highest loading in each of the factors. The values in the first column of the pair are for the multiple-group loadings while the second are for loadings from the final rotated matrix.

A	A'	B	B'	C	C'	D	D'	E	E'
91	81	10	12	34	07	57	26	30	43
22	46	88	77	02	08	46	62	-14	04
11	20	-10	-10	76	79	15	05	30	-01
38	35	06	-35	-27	03	84	82	10	-01
46	16	-13	-38	04	17	18	10	89	91

Basic and prominent general attitudes underlie the original 27×27 intercorrelation matrix as testified by the very high 72% of the total variance accounted for by the five factors. The total estimated communality was 19.12, which was completely accounted for by the five factors. Communalities, calculated from orthogonal loadings, are listed in Table 1. Intercorrelation between factors is generally high as seen in Table 2.

TABLE 1
ROTATED OBLIQUE FACTOR LOADINGS

Item Description	Factors					Σa_i^2
	Anx.- Dysph.	Alien.	Pres. Int.	Phys. Compl.	Incap.	
Always fearful*	81	12	07	26	43	87
Quite unhappy	81	45	31	40	19	72
Need much sleep*	80	-05	10	32	-08	81
Sick frequently*	79	11	27	18	22	75
Feel unloved*	78	17	02	49	49	86
Restless sleep	78	39	-01	28	36	77
Lonely	77	-17	26	22	19	86
Can't keep mind on things	76	-10	26	60	36	82
Unwanted	75	17	-08	24	60	99
Best time of life is now	74	12	22	45	28	60
Worry too much*	69	05	27	67	14	66
Have high blood pressure*	67	00	57	54	44	74
If there's a Heaven I'll go*	60	33	44	-02	03	67
Best time of life is when child*	45	12	39	-04	-06	44
Don't care what happens to me*	46	77	08	62	04	94
Family doesn't care about me*	30	74	-08	22	45	78
No self-respect left*	36	73	24	65	34	94
Don't care to see relatives	06	71	16	49	24	91
Nothing left to live for*	15	70	32	41	42	87
Don't believe in God*	07	66	03	40	46	83
Nothing interests me anymore	39	66	14	60	14	75
Don't eat enough	27	62	-04	54	-11	76
Don't think will live a year	28	59	30	58	05	70
Don't like to go visiting*	-08	29	-16	-20	-06	18
Have lots of friends	-22	03	93	13	15	115
Very religious*	20	-10	79	05	-01	74
Money root of evil	56	37	78	37	31	90
Wish had more freedom*	00	-28	75	57	52	90
Wars root of all trouble	69	08	75	42	45	92
Mentally younger than appearance	06	16	66	14	-02	55
Older cast off by younger	66	04	63	32	07	76
Trouble walking*	16	-22	60	52	59	68
Life has been tragic	27	-10	60	39	45	50
Wish had better clothes*	21	-43	59	20	06	57
People inconsiderate of others*	41	19	59	-01	-13	63

TABLE 1—(Continued)

Item Description	Factors					Σa_{ij}^2
	Anx.-Dysph.	Alien.	Pres. Int.	Phys. Compl.	Incap.	
No sex interests	37	-21	33	90	07	95
Nothing left to live for	52	51	50	84	32	96
Bad headaches frequent*	35	-35	03	82	-01	106
Fair to retire people at 65	12	-01	41	80	-17	93
Stomach trouble	55	-04	39	76	13	68
Worry about health*	72	23	36	73	02	77
Dizzy spells	49	-22	39	70	15	67
Feel helpless	63	-25	57	69	54	96
Often very unhappy	60	-13	18	68	47	79
Wish had better education	36	20	33	62	13	43
Part of body paralyzed*	16	-38	17	10	91	108
Way of living very unnatural*	02	-04	43	30	79	73
Women ruined my life	32	02	12	07	77	74
Drs. & nurses don't care about us	45	69	24	54	75	111
Have not lived good life*	11	07	-29	-05	68	70
Get annoyed easily	65	-31	13	19	64	113
Future holds nothing*	48	47	-01	30	60	69
Young can't be bothered with old	32	32	08	18	57	45
People make own troubles*	13	08	25	09	55	36
Death a relief from suffering*	06	02	-24	-05	25	17
						$\Sigma a_{ij}^2 = 19.42$

Note.—Items employed in the factor analysis are identified by asterisks. Decimal points omitted.

TABLE 2
INTERCORRELATIONS BETWEEN OBLIQUE
FACTORS

	Anx.-Dysph.	Alien.	Pres. Int.	Phys. Compl.	Incap.
Anxiety-Dysphoria		.33	.27	.46	.13
Alienation			-.05	.08	.00
Preserved Interest				.40	.18
Physical Complaints					.31
Incapacitation					

DISCUSSION

The *Anxiety-Dysphoria* factor combines the feelings of fear, tension, and being unwanted. Preoccupations with health and social situation reflect personal instability and general uneasiness. Since the manifestations are largely subjective, a superficially satisfactory adjustment is not precluded by their presence. However, the underlying lack of self-confidence and general insecurity represent a deficiency in an attribute necessary for flexible adaptation to environmental change.

The *Alienation* factor demonstrates hostility and dysphoria as a reaction to feeling rejected. It is a disgruntled reaction reflecting ambivalence toward dependency. While love and help are desired strongly, these needs are vehemently denied. Hostility, which drives others away, serves as a defense against succorance by them. These attitudes probably find ready expression in response to efforts of others to establish independence in the aged.

The *Preserved Interest* factor represents a relatively high level of interest in self and environment. While there may be some narrowing of interests with aging, and certainly reduced activity, the factor reflects strength and scope of interest as a resource. This characteristic permits a high level of social activity which may lead to the rewards of being well liked. Triteness and superficiality enter into determining this factor so that while possession of the characteristics of this factor may be essential to being interesting and well liked, garrulousness may have an adverse effect.

The *Physical Complaints* factor represents a focusing of interest on the self in terms of body functions. Since there is no control for physiologically based illness built into the study, we must assume a variety of reasons for the complaints. They may be based upon systemic malfunctioning and anatomical changes associated with aging, chronic or acute disease, or may represent a hypochondriacal exaggeration.

The *Incapacitation* factor is based upon crippling physical disease and the reaction to it. The significance of some of the heavily loaded items is not apparent and may be rather specific for the sample of subjects employed. However, it is a sizeable factor and cannot be disregarded.

SUMMARY

Geriatric attitudes of 48 elderly residents of a veterans administration center were sampled. Analysis revealed five important attitudinal factors: Anxiety-Dysphoria, Alienation, Preserved Interest, Physical Complaints, and Incapacitation.

REFERENCES

- CATTELL, R. B. *Factor analysis*. New York: Harper, 1952.
- CAVEN, R. S., BURCESS, E. W., HAVIGHURST, R. J., & GOLDHAMER, H. *Personal adjustment in old age*. Chicago: Science Research Associates, 1949.
- GUERTIN, W. H., & KRUGMAN, A. D. *The Geriatric Attitude Scale*. Gainesville, Florida: Cooperative Psychological Test Distributors, 1961.

(Received January 15, 1960)

CLINICAL JUDGMENTS AND THE DRAW-A-PERSON TEST

ROBERT E. STOLTZ AND FRANCES C. COLTHARP

Southern Methodist University

The Draw-A-Person Test (DAP), as developed by Machover (1949) and Goodenough (1926), has become an important part of the clinical psychologist's battery of assessment techniques. While widely used to provide information regarding the intellectual functioning and emotional and social behavior of a person, there is a paucity of adequate data regarding its empirical validity. Swenson (1957), in an extensive survey of the literature regarding human figure drawings, found modest confirmation of some of Machover's hypotheses regarding group trends, but little evidence of the value of the DAP for individual diagnosis. The evidence would also indicate that the validity of the method for determining the level of a person's intellectual functioning is better established than is its validity for predicting the more complex and less well defined patterns of social and emotional behavior. An additional deficiency in the existing data is that these studies which have attempted to determine the validity of the DAP for predicting social behavior have tended to use the pooled judgments of clinicians rather than investigating the problem of the individual clinician's contribution to the resultant prediction, e.g., Tolor and Tolor (1955).

This study represents an attempt to provide additional data regarding the validity of the DAP with regard to predicting intellectual, social, and emotional criteria, and also to provide data regarding the ability of individual clinicians to make such predictions.

METHOD

Subjects were 60 fourth grade school children, each of whom furnished a drawing of a person of each sex, done according to the usual DAP procedures. The judges were three clinical psychologists, each of whom was experienced in the use of the DAP and regarded as professionally competent. The judges

were asked to rate the drawings for the traits of Intelligence, Sociability, and Emotional Maturity. The ratings were done using a nine-point scale. The only information available to the judges was the age and sex of the child, which of the two drawings was completed first, and a detailed description of the testing procedure and criterion definitions for the three traits.

The criterion for Intelligence was the child's score on the Otis Quick-Scoring Mental Ability Test (Beta, Form EM), the criterion for Sociability was the preference rating given to the child by his fellow students on a sociogram, and the Emotional Maturity of the student was judged from a teacher's rating in which each teacher nominated the five most well-adjusted and the five most seriously emotionally disturbed boys and girls in her room. The teacher nomination form developed by Smith (1958) was used. On the basis of this technique the subjects were divided into groups of poor, average, and above average adjustment.

RESULTS

Distributions of the ratings for each of the traits by each of the judges were examined and found to be generally normally distributed, as were the criterion scores. The assumptions for computing the Pearson product-moment correlation were met, and this method of statistical analysis was used.¹ As the cri-

¹ All statistical computations were done with the assistance of the Southern Methodist University Computing Laboratory on the Univac 1103.

TABLE 1
CORRELATIONS BETWEEN JUDGES' RATINGS AND
INDEPENDENT MEASUREMENTS

	Intelligence	Sociability	Emotional Adjustment
Judge A	.253*	.158	-.142
Judge B	.496**	.006	.151
Judge C	.583**	.179	.165
Average for Judges A, B, C	.455**	.115	1.53

* $p < .05$.

** $p < .01$.

TABLE 2
CORRELATIONS BETWEEN JUDGMENTS
AND JUDGES

Traits	Judge		
	A	B	C
Intelligence-Sociability	.176	.337**	.609**
Intelligence-Emotional Adjustment	.381**	.447**	.668**
Sociability-Emotional Adjustment	.082	.872**	.881**
Average Intercorrelations	.220	.615**	.750**

** $p < .01$.

terion for emotional adjustment was essentially trichotomous, all correlations regarding this trait were corrected for coarseness of grouping.

Table 1 gives the correlations between the judges' ratings and the criterion for each of the three traits, as well as the average judge intercorrelations for each trait.² Table 2 gives the intercorrelations between the ratings for the three traits by each of the judges, as well as the average trait intercorrelation for each judge. Table 3 gives the intercorrelations between the three judges for each of the three traits as well as the average intercorrelations of the judges for each of the traits.

DISCUSSION

As is evident in Table 1, the three judges were able to predict the intelligence test performance of the subjects to a significant degree. However, there is a considerable difference among the judges in their ability to predict this criterion. For example, a test of the significance of a difference between correlation coefficients indicated that Judge C

² All correlations were converted to Fisher z coefficients for averaging and then converted back to correlation coefficients.

is significantly better able to predict this trait than is Judge A. None of the judges is able to predict the Sociability criterion or the Emotional Adjustment criterion to a significant degree. Even more distressing is Judge A's prediction of Emotional Adjustment which is negatively correlated with the criterion. Efforts to obtain multiple correlations between optimally combined judges' ratings and the trait criteria resulted in multiple R s of .614, .230, and .293 for Intelligence, Sociability, and Emotional Adjustment, respectively. We must conclude that even optimum weighting of the clinician's judgments does not produce significant prediction of the latter two criteria.

In the design of the study an effort was made to choose criterion areas which would be distinguishable from each other, and whose specific criterion scores would tend to be independent of each other. The intercorrelations of the criteria furnish some data bearing on the extent to which this effort was successful. The correlation between the intelligence scores and the sociogram scores was .414; between intelligence scores and the teacher ratings on Emotional Adjustment, .495; and between the sociogram scores and the teachers' ratings of Emotional Adjustment, .458. Each of these correlations is significant beyond the .01 level. It can be concluded that the criteria were not completely independent, but were only moderately so.

From Table 1 it can be concluded that Judge C is the best judge of the criteria, while Judge A seems to be the poorest overall. This difference in the ability to predict the criteria seems to be due to the joint influence of a judge's ability to predict a subject's intelligence test performance and to the extent

TABLE 3
CORRELATIONS BETWEEN JUDGES

Judges	Intelligence	Sociability	Emotional Adjustment	Average Correlation for Judges
A-B	.579**	.243	.307*	.390
A-C	.555**	.321*	.349**	.445
B-C	.702**	.495**	.504**	.657
Average on Variables	.615**	.350**	.390**	

* $p < .02$.
** $p < .01$.

to which a given judge implicitly viewed intelligence as a trait related to Sociability and Emotional Adjustment. As indicated above, both the Sociability and Emotional Adjustment criteria were correlated with the Intelligence criterion. Partialing out the effect of the Intelligence criterion reduced the correlation between Sociability and Emotional Adjustment to .320, indicating a fairly large contribution to the zero order correlation between these two criteria. Table 2 shows quite clearly that Judge C produced trait ratings that were highly correlated with each other, while Judge A's trait ratings showed a significant correlation only in the case of the Intelligence-Emotional Adjustment traits. Since the Intelligence criterion contributes heavily to the other criteria, and since Judge C is not only the best predictor of this criterion, but also the judge who shows the strongest tendency to use ratings on this dimension as an element in predicting the other criteria, it would follow that Judge C would appear to be the most accurate judge overall. The converse of this argument would hold for Judge A, and the same argument would support the appearance of Judge B as the judge with an intermediate degree of overall predictive ability. It would appear that the impression of overall predictive accuracy conveyed by a given clinician is, in this case at least, a function of the clinician's ability to develop a valid index of Intelligence from the DAP and little more.

Table 3 reflects the extent to which the judges agree with each other in their ratings of the three traits. As would be expected from the other tables, the judges show their best agreement when it comes to predicting Intelligence from the DAP, and much less agreement when it comes to predicting the other two criteria. The higher correlations between Judges B and C can be explained by the greater extent to which the judges appear to be reacting to a general factor in their trait definitions, and would indicate that these general factors are similar for the two judges. The present study was not designed to tell us just what this general factor may be. It may be simply a liking for some drawings over others, a reliance on certain common references in the literature, a carry-over from their general impression of the person's ad-

justment, or it may be some halo effect derived from a source in the drawings as yet unknown.

The findings of this study would not, in general, support the use of the DAP as a measure for predicting behavior criteria in the area of Sociability or Emotional Adjustment. The findings would lend support to the use of the DAP as a measure of intellectual functioning, but would also support the earlier reviews which suggest that the relationships are not adequate for individual prediction.

SUMMARY

The present study was designed to provide additional data regarding the ability of clinical psychologists to predict criteria of intelligence, sociability, and emotional adjustment from human figure drawings. The subjects were 60 fourth grade school children who were given the Draw-A-Person Test in a group situation. Three clinical psychologists judged the extent to which each child's drawings indicate the existence of one of the three trait criteria. The relations of the clinical judgments to the criteria were statistically compared.

The psychologists were able to predict intelligence to a statistically significant degree, but were unable to predict either sociability or emotional adjustment. Although working independently, the judges did show a significant amount of correlation with each other in their predictions.

Factors influencing the ability of the judges to produce ratings that would correlate with the criteria are discussed.

REFERENCES

- GOODENOUGH, FLORENCE L. *Measurement of intelligence by drawings*. Yonkers-on-Hudson: World Book, 1926.
- MACHOVER, KAREN. *Personality projection in the drawings of a human figure*. Springfield, Ill.: Charles C Thomas, 1949.
- SMITH, L. M. The concurrent validity of six personality and adjustment tests for children. *Psychol. Monogr.*, 1958, 72(4, Whole No. 457).
- SWENSON, C. H., JR. Empirical evaluations of human figure drawings. *Psychol. Bull.*, 1957, 54, 431-466.
- TOLOR, A., & TOLOR, BELLE. Judgment of children's popularity from their figure drawings. *J. proj. Tech.*, 1955, 19, 170-176.

(Received January 22, 1960)

AN APPLICATION OF PREDICTION TABLES TO THE STUDY OF DELINQUENCY¹

PETER F. BRIGGS, ROBERT D. WIRT, AND ROCHELLE JOHNSON

University of Minnesota

The rate of occurrence of a characteristic in a specified population has come to be called the *base rate* of the characteristic. Meehl and Rosen (1955) have discussed the importance of considering base rates in evaluating a predictive system. They point out that "a psychometric device, to be efficient, must make possible a greater number of correct decisions than could be made in terms of the base rates alone"² (p. 194). An illustration used by these authors was the prediction of juvenile delinquency by the Gluecks (Glueck & Glueck, 1950). In their example where the base rate concept had been ignored, the data were in effect treated as though the base rate were 50%, which is highly unlikely. The present authors have cross-validated the same predictors and the conclusions drawn by Meehl and Rosen were born out (Wirt & Briggs, 1960).

Further examples of the importance of the base rate can be found with ease. An interesting recent article by Schofield and Balian (1959) compares the incidence of psychic trauma among normal and schizophrenic patients. Their results indicate that the base rate for trauma is so high that it is obviously not peculiar to their schizophrenic sample. This study followed the form suggested by Pearson and Kley (1957) who remark:

Eventually, like it or no, we will have to come to grips with the high probability that the base rate

problem applies in the prediction of mental disorder from kind and number of traumatic life experiences, just as it applies in the case of psychometric prediction (p. 407).

Pearson and Kley go on to lament with others the fact that behavioral scientists do not tend to consider base rates in their study of case materials.

The present study suggests that prediction in the area of delinquency is not an altogether lost cause at this time if one's goals are reasonable or moderate and one tends to remain in the relatively narrow context in which the criterion data were gathered. In this connection Pearson and Kley (1957) suggest

... that individuals in a population with a known and relatively high incidence rate for a particular disorder may be submitted to longitudinal investigation of a kind, which would not be economical for samples drawn from the general population (p. 400).

Although their argument was aimed at the discovery of etiological factors, one may also examine efficiency of a treatment program through the study of a highly concentrated sample of cases among whom pathology may be expected to occur.

The approach used here was the multiple criteria technique discussed by Meehl and Rosen (1955). The first criterion was the Minnesota Multiphasic Personality Inventory (MMPI); cases were selected whose MMPI profile had certain scale elevations that were known to be related to delinquency (Hathaway & Monachesi, 1951). The second criterion was selection *within* this group using family history data developed in another study by the authors. It was shown in the earlier study that a great number of family history factors, especially those commonly recognized as tragic, can be related to delinquency (Wirt & Briggs, 1959). Thus with

¹ This research was supported in part by Grant 1151c from the National Institute of Mental Health, Public Health Service, United States Department of Health, Education and Welfare; and in part by the Graduate School of the University of Minnesota. The authors wish to express their appreciation to the federal government and to their university.

² There is one minor exception to this rule, viz., when the valid positive rate equals the valid negative rate, accuracy of prediction is independent of the base rates.

a well recognized personality technique and one of a number of possible indicators of family disorder, it was possible to obtain the results demonstrated below.

METHOD

Samples

From a previously described sample of nearly two thousand boys tested by Hathaway and Monachesi (1951) during the school year 1947-48, a sample of 573 cases stratified on MMPI code and delinquency was drawn from a population of 1,958 cases. These subjects had been tested in the ninth grade and most of them were thirteen years old at the time. Delinquency ratings were based on the period following testing, so that in this study the *prediction* of delinquency refers to the prediction of a subsequent phenomenon.

The sample of boys was dichotomized, one group was composed of boys whose MMPI codes contained combinations of three delinquency "excitor" scales and this subsample of 201 cases represented 550 and this subsample of 201 cases represented 550 cases in the population. The remaining group was composed of boys whose MMPI codes did not include the excitor combinations. There were 372 cases representing 1,408 cases in the population.³ The excitor MMPI scale combinations were: *Pd-Sc*, *Pd-Ma*, *Sc-Pd*, *Sc-Ma*, *Ma-Pd*, and *Ma-Sc*. (See Hathaway & Monachesi, 1951, and Wirt & Briggs, 1959, for justification of these procedures.) Since the code numbers for these scales are 4 = *Pd*, 8 = *Sc*, and 9 = *Ma*, the excitor code sample is called the "489" group.

Most studies of delinquents have used some contact with the police as a defining criterion, omitting the question of severity altogether. Delinquent severity is not a homogeneous personality dimension but probably reflects a sociological or judgmental dimension of the society or of the perceiver of the delinquent (Wirt & Briggs, 1959). The definition of delinquency adhered to in this study was a rating delinquency adhered to in this study was a rating based upon court and police docket records. A dichotomous split was made: (a) a less severe criterion including all cases who had any contact with the police and (b) a more severe criterion excluding persons whose contacts with the police involved only minor infractions.

Family History Factors

The source from which these data were derived was a survey of the records of 11 social agencies (voluntary and governmental). The case records of all boys and their families which were identified in the admission files of the 11 agencies were completely reviewed for data which seemed psychologically important. The data were collected in note

³ The sample/population proportions are not equitable because the samples actually include subsamples which were weighted differentially to produce a valid estimate of population values.

form and were found to include 42 fairly common but discrete items. These items were grouped into seven more general categories: family disruption, poverty or need, dissocial behavior, psychiatry for family, marital disruption, inadequate parent-child relationship, and minor psychological problems. The present study focused on data from one of these categories, "family disruptions due to disease," which included six items: mother dies, father dies, mother chronically ill, father chronically ill, siblings die, or siblings are chronically ill. A category score could be determined by assigning one point for each item present in the family history. Thus a score of 1 point meant that at least one of the items was true for a given family, 2 points meant that two items were true or that one item occurred twice, etc.

It is to be understood that these items of information were recorded in the social agency records *before* the subjects had been delinquent and *before* they were tested by Hathaway and Monachesi in 1948. The delinquency that is referred to occurred after testing and thus also after any particular disruptions of the family due to disease.

RESULTS

The data are presented for two degrees of delinquency. The less severe criterion, which nets of course the largest percentage of delinquents, shows the estimated overall population rate of delinquency to be 41%. Among cases with elevated excitor (i.e., 489) codes, the proportion of delinquents was approximately 43%. Using the slightly more severe criterion of delinquency in which minor offenses were excluded from the delinquent sample, the overall population rate was 32% delinquent, while the rate for the 489s was 35% delinquent.

Among cases with instances of family disruption due to disease, the rate of delinquency for all codes tended to increase as the number of family disruptions due to disease increased. That is, accuracy of delinquency prediction improved for cases known to have family disruptions due to disease regardless of MMPI patterns of the subjects. By calling the whole population delinquent the accuracy of prediction would be only 32%; by calling cases with a score of 1 or more points on family disruption due to disease delinquent, accuracy of prediction would rise to 43%, at a score of 2 or more points accuracy would reach 53%, and accuracy of such prediction increases to a maximum of 63%. Thus if one is interested in selecting potential delinquents for treatment, knowledge of social agency con-

TABLE 1

PROPORTION OF BOYS WITH 489 MMPI CODES AT EACH OF SEVEN LEVELS OF FAMILY DISRUPTION DUE TO DISEASE WHO BECAME DELINQUENT AT A BASE RATE OF 43% DELINQUENT PREDICTION RATIOS

Number of Disruptions	Delinquents		Nondelinquents			Pp_1	Qp_2	Qq_2	H_T	Rp	Hp
	Est. cf	p_1	Est. cf	p_2	q_2						
0	235	1.00	315	1.00	.00	.43	.57	.00	.43	1.00	.43
1	57	.24	51	.16	.84	.10	.09	.48	.58	.20	.53
2	35	.15	16	.05	.95	.06	.03	.54	.61	.09	.70
3	27	.12	4	.01	.99	.05	.01	.57	.61	.06	.88
4	16	.07	4	.01	.99	.03	.01	.57	.60	.04	.81
5	15	.06	4	.01	.99	.03	.01	.57	.59	.03	.79
6	8	.03	4	.01	.99	.02	.01	.57	.58	.02	.68
7+	4	.02	00	.00	1.00	.01	.00	.57	.58	.01	1.00

Note.—Apparent discrepancies in table are due to rounding figures.

tact would be an asset. It should be noted, however, that this does not aid much in prediction of nondelinquency.

When MMPI code is taken into account, accuracy in prediction of delinquency is increased even more. The cases with 489 codes have a frequency of 28% in the population. The complete results for the 489 cases are presented in Tables 1 and 2 for the two different levels of delinquent severity. This is a demonstration of the prediction model developed by Meehl and Rosen (1955). Here the first column indicates the score (i.e., the number of disruptions in the family due to disease) from 0 through 7 or more. The second column gives the cumulative frequencies of these 489 cases who became delinquent at

each social agency score from a maximum of 7 or more points to a minimum of 0; these are estimates of the population cumulative frequencies. The column entitled p_1 transforms these cumulative frequencies to proportions based on the total number of 489 cases that became delinquent. The third column gives the cumulative frequencies for those 489 cases that did not become delinquent at each social agency score. Correspondingly, the column p_2 transforms these cumulative frequencies to proportions based on the total number of 489 nondelinquents.

For example, a p_1 of .17 at a family disruption score of 2 indicates that 17% of the 489 cases who became delinquent scored 2 or more points on this particular social agency

TABLE 2

PROPORTION OF BOYS WITH 489 MMPI CODES AT EACH OF SEVEN LEVELS OF FAMILY DISRUPTION DUE TO DISEASE WHO BECAME DELINQUENT AT A BASE RATE OF 35% DELINQUENT PREDICTION RATIOS

Number of Disruptions	Delinquents		Nondelinquents			Pp_1	Qp_2	Qq_2	H_T	Rp	Hp
	Est. cf	p_1	Est. cf	p_2	q_2						
0	191	1.00	359	1.00	.00	.35	.65	1.00	.35	1.00	.35
1	45	.24	63	.18	.82	.08	.11	.54	.62	.20	.42
2	33	.17	18	.05	.95	.06	.03	.62	.68	.09	.65
3	25	.13	7	.02	.98	.05	.01	.64	.68	.06	.79
4	15	.08	5	.01	.98	.03	.01	.64	.67	.04	.73
5	15	.08	4	.01	.99	.03	.01	.64	.67	.03	.79
6	8	.04	4	.01	.99	.02	.01	.64	.66	.02	.68
7+	4	.02	0	.00	1.00	.01	.00	.65	.66	.01	1.00

Note.—Apparent discrepancies in table are due to rounding figures.

scale. Similarly, a p_2 of .05 indicates that only 5% of the nondelinquents scored 2 or more points on this scale.

The proportions given in p_1 and p_2 do not take into account the base rates.

Column q_2 represents the valid negative rate before the base rates are taken into account. This indicates the proportion of all 489 nondelinquents that are appropriately labeled "nondelinquent" at each social agency scale score. The column entitled Pp_1 is P_1 multiplied by the base rate of delinquency for 489s and represents the valid positive rate. The column Qp_2 is p_2 multiplied by the base rate of nondelinquency and represents the valid false positive rate. Qq_2 is q_2 multiplied by the base rate of nondelinquency and represents the valid negative rate. H_T is $Pp_1 + Qq_2$ and represents the overall accuracy of prediction, which is the accuracy with which one can call individuals falling above the given cutting scores nondelinquents and those at cutting scores delinquents. The column called Rp is $Pp_1 + Qp_2$ and represents a sort of selection ratio telling the proportion of people for whom one is predicting at each level on the social agency rating (i.e., at a score of 1 or greater, a score of 2 or greater, etc.). And finally, a column Hp is Pp_1/Rp and it represents the proportion of people at each level on the social agency rating who are delinquent. This is the accuracy with which one can predict delinquency alone with no regard for false negatives. It is the most useful statistic if one is trying to select a small sample for treatment.

These tables show the advantage of proceeding from a high base rate of delinquency. The selection ratio, Rp , indicating the percentage of people for whom prediction is made at each level does not change appreciably with the change in rate (compare Rp in Table 1 with Rp in Table 2). Yet the accuracy of delinquency prediction, Hp , (and therefore the percentage of people at each level who are delinquent and who are correctly identified) increases as the percentage of delinquents increase in the population. It should be noted that at the more severe criteria, with a 35% delinquency rate in the overall population, it is possible to identify a sample that is 79% delinquent although

TABLE 3
NUMBER OF CASES WHICH WOULD BE FOUND DELIN-
QUENT FOR THREE TYPES OF SELECTION STARTING
WITH 1,000 BOYS CHOSEN RANDOMLY

	Severity	Success at Each Rate		
		N Selected	% Del.	N Del.
Random	Severe	1000	32	316
	Mild	1000	41	408
All 489	Severe	281	35	98
	Mild	281	43	120
489 and	Severe	16	79	13
≥ 3 Disruptions	Mild	16	88	14

this group represents only approximately 6% of the 489 cases. At the less severe criteria of delinquency where the overall rate is 43%, it is possible to identify a sample that is almost 88% delinquent and again it includes only 6% of the cases in the 489 population.

The success with which delinquency can be predicted based on a hypothetical 1,000 cases using the information described above is shown in Table 3. Depending upon the needs of the individual situation, it is possible to predict that 316 boys in a population of 1,000 would be fairly severely delinquent and as useful information is added, one can follow the 1,000 cases down to the precise but restricted sample obtained through the study of 16 cases of whom 14 will be delinquent.

DISCUSSION

The principles involved in selection described here are empirical rather than theoretical and do not carry general implications concerning the cases selected. It would be judged from other data not presented that a number of such techniques could be developed using different factors within the histories of delinquent boys. Some such techniques would probably be better than others and some would be more stable than others. Without theoretical knowledge of the reasons for the operation of each of the particular criteria established, it would be impossible to know when environmental or cultural changes would affect the validity of the techniques that are presented or developed. These tend to be the risks involved in using such an approach as

the present one. Of course, it is not impossible to build in safeguards, to define populations, and to cross-validate from time to time, in order to raise the likelihood that one is not practicing an impoverished statistical ritual.

It was indicated earlier that it is likely that delinquency is not a homogeneous psychological variable. It is obvious that the selection of delinquents through some particular set of criteria will net a very special subpopulation of delinquents, a population which is not a random sample of all delinquents. Therefore, the research worker employing selection techniques based upon factors which heighten the likelihood of delinquency in a specific way will tend to meet with a specific type of case. This should be an advantage since there is some insurance of homogeneity within the subpopulation of delinquency besides the likelihood of future delinquency. Furthermore, the procedures themselves, although not theoretically involved in the understanding of delinquency, certainly suggest some factors which are important in the development of delinquency within the cases selected. Therefore, within the framework of criteria described in this paper, where delinquency seems to result from a combination of present personality status (characterized by poor judgment, excitability, and a certain unrealistic approach to the events of life) plus an abnormal home in which tragedy and disease have left their permanent marks, suggestions for treatment are not as hard to make as if one were dealing with the random delinquent. Furthermore, one would suspect that cases selected in the same way might require similar treatment programs since such selection is a quasidiagnostic procedure.

SUMMARY

A technique for the discovery and identification of potentially delinquent boys was illustrated in this paper. A sample of 13-year-old boys, drawn randomly from a general urban population, was evaluated using the MMPI and a survey of their family histories

obtained from social agencies. A multiple criteria approach to the identification of the pre-delinquent case was developed, starting with cases from the general population. To this the factors of MMPI codes and instances of severe disease or of death among members of the family were added. Using these two criteria, it was possible to develop small subpopulations which were about 80% saturated with pre-delinquent boys. Such subsamples when compared with the general population were approximately twice as dense with pre-delinquent cases since the general population had a rate of about 40%.

The possibility of using this technique in the establishment of treatment programs where small samples could be handled in areas where large numbers of delinquents are found was discussed. It was pointed out that such subpopulations would not be random samples of delinquents, but that such subpopulations would in fact be more homogeneous subsamples of delinquents than are usually obtained.

REFERENCES

- GLUECK, S., & GLUECK, ELEANOR. *Unravelling juvenile delinquency*. New York: Commonwealth Fund, 1950.
- HATHAWAY, S. R., & MONACHESI, E. D. *Analyzing and predicting juvenile delinquency with the MMPI*. Minneapolis: Univer. Minnesota Press, 1951.
- MEEHL, P. E., & ROSEN, A. Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychol. Bull.*, 1955, 52, 194-216.
- PEARSON, J. S., & KLEY, IRENE B. On the application of genetic expectancies as age-specific base rates in the study of human behavior disorders. *Psychol. Bull.*, 1957, 54, 406-420.
- SCHOFIELD, W., & BALIAN, LUCY. A comparative study of the personal histories of schizophrenic and nonpsychiatric patients. *J. abnorm. soc. Psychol.*, 1959, 59, 216-225.
- WIRT, R. D., & BRIGGS, P. F. Personality and environmental factors in the development of delinquency. *Psychol. Monogr.*, 1959, 73(15, Whole No. 485).
- WIRT, R. D., & BRIGGS, P. F. The efficacy of ten of the Glueck's predictors. *J. crim. Law Criminol. police Sci.*, 1960, 50, 478-479.

(Received January 25, 1960)

SEXUAL IDENTIFICATION AND THE FIRST FIGURE DRAWN

RENATE G. ARMSTRONG AND PAUL A. HAUCK¹

East Moline State Hospital, Illinois

Statements regarding interpretation of projective tests, such as the House-Tree-Person or the Figure Drawing Tests, were, in good part, originally based on intuition. To date, little experimental evidence has come forth to support or deny many of the assertions made by clinicians using such instruments.

A frequently accepted hypothesis regarding the Draw-A-Person Test is that a subject (S) who draws first a person of the opposite sex, has a problem of sexual identification. While Machover (1949) has adduced empirical evidence in support of the hypothesis, Granick and Smith (1953), Barker, Mathis, and Powers (1953), and several other studies cited in Swenson's review article (1957) failed to confirm the asserted relationship.

The present study relates the order of future drawing to the discrepancy between self-concept and ideal self, mother, and father—as determined from the Leary Interpersonal Check List (1956). Four specific hypotheses were tested:

1. Ss drawing the opposite sex first have different self-concepts (as revealed by the Leary Interpersonal Check List) from Ss drawing the same sex first.

2. Ss drawing the opposite sex first have a greater similarity of self-concept with their perception of the parent of the opposite sex than Ss drawing the same sex first.

3. Ss drawing the opposite sex first will show a greater discrepancy of the self-concept with their ideal self-concept than Ss drawing the same sex first.

4. If Hypothesis 3 is supported, then those Ss who draw the opposite sex first will have an ideal self-concept resembling that of the parent of the same sex more than Ss who draw the same sex first.

METHOD

One hundred and fourteen undergraduate college students (57 females, 57 males) were given the Draw-A-Person Test and the Leary Interpersonal Check List (1956). The mean age for the females was 20.54 years and for the males almost 21.77 years. This difference in age is significant at the .02 level. Although our male and female samples differ significantly in age, it is felt that this is not a serious drawback, particularly in view of Swenson and Newton's (1955) study which indicated that age differences were not a significant factor in sex differentiation beyond the eighth grade. The mean years of college education for the females was 2.93 and for the males 2.58, while for the two combined it was 2.75. The difference in years of education is not significant ($t = 1.46$).

On the Leary each S was instructed to agree or disagree with 128 descriptive words or phrases which he would use in describing himself, his mother, his father, and his ideal self. When performing the Draw-A-Person Test, careful note was made of the sex of the first drawn figure. This was to be the basis upon which the groups would be divided. Thirteen of the 57 females drew a male figure first, while 16 males drew the female figure first. The difference is not significant ($\chi^2 = .40$).

The template scoring system was used for each Leary protocol. Then each Dominance and Love vector was computed. The means and sigmas of the vectors were computed for the various groups. Table 1 gives these data. These groups include females drawing their own sex first (Ff), those drawing the opposite sex first (Fm), males drawing a male figure first (Mm), and the male group drawing the female first (Mf). Several t tests were run for Groups Ff and Fm, Mm and Mf, total female and total male. As a final analysis of the data Table 2 indicates the t values between the means of different

¹ Our thanks are given to Martin S. Sloane, Superintendent, for his permission to carry on this research.

TABLE 1
MEANS AND SIGMAS OF VECTORS FOR FOUR RATINGS OF ALL GROUPS

Group	Self				Mother				Father				Ideal Self			
	Dominance		Love		Dominance		Love		Dominance		Love		Dominance		Love	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Ff	54.30	8.97	51.61	6.10	61.25	7.58	54.98	7.65	64.11	8.36	49.09	9.12	65.43	3.73	56.13	5.42
Fm	57.84	6.05	53.84	5.91	61.46	6.29	59.38	6.57	61.38	5.85	47.92	9.79	67.15	6.97	51.85	6.37
F total	55.10	8.58	50.14	6.14	61.29	7.38	55.98	7.67	64.17	7.87	48.42	9.30	65.82	4.76	56.07	5.68
Mm	57.54	7.40	48.85	8.57	61.22	6.50	55.68	7.82	62.29	8.59	49.98	8.92	61.95	6.06	51.85	5.15
Mf	57.62	5.98	50.19	6.33	59.37	7.31	54.75	8.24	63.19	5.55	47.06	12.85	63.62	5.96	55.06	7.67
M total	57.56	7.15	49.23	7.75	60.70	6.80	55.42	7.94	62.54	7.87	49.16	10.26	64.58	6.02	54.91	5.96
Total M & F	56.14	9.21	50.68	7.32	60.99	7.15	55.55	8.81	63.35	7.96	48.99	9.80	65.23	5.09	55.49	5.85

ratings on Dominance and Love vectors for all groups.

RESULTS

Few differences between group means were significant. Out of 24 *t*'s, only two reached the 5% level: men vs. women, on self-concepts scored for Love, and women drawing same sex first vs. women drawing opposite sex first, on mother concept scored for Love. These last findings could well be chance fluctuations.

The groups as a whole (considered on the basis of their sex differences only) showed a difference between the means of the Love vectors ($p = .05$) on the self-rating scale. Examination of Table 1 shows us that the females rate themselves higher on those dimensions tapped by the Love vector, while males scored higher on the Dominance vector. The implications of this will be dealt with in the discussion section.

In Table 2 we can examine the significance tests between the means of the various ratings for the Dominance and Love vectors for

the Ff, Fm, Mm, and Mf groups. The Ff row shows us that the only difference which is not significant for the Dominance vector is that between Father and Ideal Self. Two differences fail to reach significance, however, for the Love vector, Self-Father, and Mother-Ideal Self. The Fm group shows a different pattern from the Ff in that the Self-Mother ratings on the Dominance vector are not significantly different. On the other ratings in this vector these two groups are essentially similar. For the Love vector the Fm group is quite similar on most of the ratings with the exception of the Self-Ideal Self ratings. The Fm sample does not differ on these two ratings, while for the Ff sample the difference between Self and Ideal Self is highly significant ($p = .001$).

The males who drew the male figures first (Mm) have significant differences between all the ratings on the Dominance vector except Father-Ideal Self, a finding common to Ff and Fm groups as well. Two comparisons of ratings have small differences on the Love

TABLE 2
t's BASED ON MEAN DIFFERENCES BETWEEN INDICATED PAIRS OF RATINGS FOR VARIOUS GROUPS

Group	Dominance Vector					Love Vector				
	Self-mother	Self-father	Self-Ideal Self	Mother-Ideal Self	Father-Ideal Self	Self-mother	Self-father	Self-Ideal Self	Mother-Ideal Self	Father-Ideal Self
Ff	-6.95**	-9.81**	-9.13**	-4.18**	-1.32	-3.34*	+2.55	-5.79**	-1.45	-7.34**
Fm	-3.62	-6.54*	-9.31**	-5.69*	-2.77	-5.54*	+5.92	-1.01	+4.53	-6.93*
F total	-6.19**	-9.07**	-10.75**	-4.53**	-1.65	-3.84**	+3.32*	-3.93**	-0.09	-7.25**
Mm	-3.68*	-4.75**	-7.41**	-3.73**	-2.66	-6.83**	-1.13	-6.00**	+0.83	-4.87**
Mf	-1.75	-5.57*	-6.00**	-4.25	-0.43	-4.56	+3.13	-4.87	-0.31	-8.00**
M total	-3.14*	-4.98**	-7.02**	-3.88**	-2.04	-6.19**	+0.07	-5.68**	+0.51	-5.75**
M & F total	-4.85**	-7.21**	-9.09**	-4.34**	-1.88*	-4.87**	+1.79	-4.81**	+0.06	-6.50**

* Significant at .05 level.

** Significant at .01 level.

vector for the Mm group. The Self-Father ratings show no significant differences, as do the Mother-Ideal Self ratings. The overall pattern of significant differences is largely similar for both groups drawing their own sex first in both the Dominance and Love vectors.

Those males drawing the female first (Mf) had no significant difference between ratings of Self and Mother on the Dominance vector, much as the Fm group also showed. And once again, as each group thus far has shown, there is no significant difference between Father and Ideal Self.

Four ratings show no significant differences in the Love vector for the Mf group. The Self-Mother, Self-Father, Self-Ideal Self, and Mother-Ideal Self ratings are without the significant differences which characterize the Mm males. The most outstanding differences between the two male groups on the Love vector are on the ratings between Self and Mother and between Self and Ideal Self, with the Mm males showing the greatest differences.

Again the overall pattern of significant differences is practically identical for both groups drawing the opposite sex first, a finding which applied to both vectors. It would appear that the sex of the S doing the drawing does not differentiate the interrater comparisons as much as the sex order in which the figures are drawn.

An inspection of the column of the comparisons between parent ratings and the Ideal Self ratings reveals a consistent trend. In the Dominance vector it is noted that all groups tend to identify (i.e., show no significant differences between Father and Ideal Self ratings) with the perceived masculine qualities of the father but not with the mother except in one instance where the Mf group does show some degree of identification between Father and Ideal Self. A similar pattern of significance tests was obtained on the Love vector except that in this case the groups unanimously wanted to identify with the Love qualities perceived in their mothers and not in their fathers.

DISCUSSION

Our study tends to support the statements made by Machover (1949). On the other

hand, our results are not in line with those studies suggesting that certain conclusions cannot be drawn from a sample which draws its own sex first as against a sample which draws the opposite sex first. We find marked differences between these two groups. These differences occur in their perception of their self and ideal self as compared to their ratings of their parents. These are not gross differences but are, rather, of a selective nature. Some appear only in relation to qualities of dominance, others to love.

We are in a position now to examine our original hypotheses. Let us repeat each and see if it is upheld or denied.

Hypothesis 1. Ss drawing the opposite sex first have different self-concepts (as revealed by the Leary Interpersonal Check List) from Ss drawing the same sex first.

This was not supported. The self-concept is measured by the Dominance and Love vectors, and the means on these vectors were not dissimilar enough to support this hypothesis, although when the sexes were combined, the *t* test done between the mean female Love score and the mean male Love score differed at the .05 level. This could be indicative of a sex difference, although it may be attributable to the larger *N*.

Hypothesis 2. Ss drawing the opposite sex first have greater similarity of self-concept with their conception of the parent of the opposite sex than Ss drawing the same sex first.

The Fm group rates the self more like the father on dominance qualities than does the Mf group Ss who also rate their self more like their mothers on dominance qualities than does the Mm group. The hypothesis is, therefore, upheld at least for Dominance vectors.

On the Love vector the hypothesis is only partially upheld. Ff and Fm groups are not different in how they perceive their self from their fathers. But there is a significant difference between how the men rate themselves and their mothers. The Mf group sees its self having love qualities like their mothers' love qualities, but the Mm group is significantly different in this respect.

Hypothesis 3. Ss drawing the opposite sex first will show a greater discrepancy of the self-concept with their ideal self-concept than Ss drawing the same sex first.

This is not confirmed. Although on the Dominance vector males and females both show a significant difference between their self and ideal self-concepts, this difference is by far more pronounced for Ff and Mm groups. On the Love vector the disagreement between self and ideal is in fact seen only for Ff and Mm groups. This could suggest that Ff and Mm groups have higher aspirations and are more ambitious for themselves, thereby not being easily satisfied. Or it could mean that drawing the opposite sex first reflects a healthier acceptance of one's goals by either being more capable of attaining the ideal, or bringing the ideal down more in keeping with the way one finds himself. This would certainly be a worthwhile subject for further research, especially in view of the work of Butler and Haigh (1954) in which therapeutic progress was found to be a decrease in discrepancy between the self and ideal self-concepts.

Hypothesis 4. This hypothesis was contingent upon the confirmation of Hypothesis 3 and is hence rejected also.

The hypotheses originally proposed have failed to deal with all of the resulting data. A few additional comments are in order.

The most outstanding finding from this study reveals that Ss of either sex, if they draw their own sex first, will tend to be basically similar regarding perceptions of self-ideal self, and parents. This also tends to be true for all Ss drawing the opposite sex first, in that they too agree with each other on these ratings.

Although most of the mean difference comparisons yielded significant *t*'s, none resulted for the Father-Ideal Self comparisons on the Dominance vector for all four groups, suggesting that all Ss ideally want to be like the dominant qualities they perceive in their fathers. On the Love vector the Self-Father and the Mother-Ideal Self comparisons failed to reach significance in any of the groups, implying that all Ss attribute similar love qualities to themselves and their fathers, although ideally they would like to have the love qualities perceived in their mothers.

We conclude that Machover's contention

has at least some inferential support from this study.

SUMMARY

One hundred and fourteen college undergraduate subjects, 57 males and 57 females, with an average college education of 2.5 years were given the Draw-A-Person Test and the Leary Interpersonal Check List. The sex of the first drawn figure was noted and the ratings of the subjects for themselves, mothers, fathers, and ideal selves were scored and compared according to Leary's Dominance and Love vectors.

An analysis of the data suggests that differences exist between the groups drawing their own sex first from groups drawing the opposite sex first. Four hypotheses were tested and research suggestions made from these data. Of these four, three were rejected and only one was partially upheld. However, the pattern of differences between self-concept and concepts of ideal, of father, and of mother in the various groups was such as to provide some inferential support for Machover's position.

REFERENCES

- BARKER, A. J., MATHIS, J. K., & POWERS, C. Drawing characteristics of male homosexuals. *J. clin. Psychol.*, 1953, 9, 185-188.
- BUTLER, J. M., & HAIGH, G. V. Changes in the relation between self-concepts and ideal concepts consequent upon client-centered counseling. In C. R. Rogers & Rosalind F. Dymond (Eds.), *Psychotherapy and personality change: Co-ordinated studies in the client-centered approach*. Chicago: Univer. Chicago Press, 1954. Pp. 55-76.
- GRANICK, S., & SMITH, L. J. Sex sequence in the Draw-A-Person Test and its relation to the MMPI Masculinity-Femininity scale. *J. consult. Psychol.*, 1953, 17, 71-73.
- LEARY, T. *Multilevel measurement of interpersonal behavior*. Berkeley, California: Psychological Consultation Service, 1956.
- MACHOVER, K. *Personality projection in the drawing of a human figure*. Springfield, Ill.: Charles C Thomas, 1949.
- SWENSON, C. H. Empirical evaluations of human figure drawings. *Psychol. Bull.*, 1957, 6, 431-466.
- SWENSON, C. H., & NEWTON, K. R. The development of sexual differentiation on the Draw-A-Person Test. *J. clin. Psychol.*, 1955, 11, 417-419.

(Received January 27, 1960)

REPLICATED FACTORS ON THE MMPI WITH FEMALE NP PATIENTS

WILLIAM J. EICHMAN

Veterans Administration Hospital, Roanoke, Virginia

Although factor analytic techniques have been applied to Minnesota Multiphasic Personality Inventory (MMPI) scales on a number of occasions, there is no study in the literature which employs female NP patients as a subject group. Consequently, it is unknown whether the factorial structure of the MMPI with female patients is different from that with male patients. Studies with male subjects indicate essential agreement as to the loadings on the first two factors although the interpretations of the factors differ from one study to the next. Welsh (1956) seems to have conceptualized the two dimensions most adequately as anxiety and repression. He developed item scales for the two factors and labeled them *A* and *R*. More than two factors have been found in all reported studies but the loadings have differed from one study to the next. Welsh (1956) identified two further factors in his study and developed item scales for them. These scales were difficult to interpret and have received little further attention. Wheeler, Little, and Lehner (1951) found "paranoid adjustment" and "psychopathic adjustment" factors. Kassebaum, Couch, and Slater (1959) found a third factor which they labeled "tender minded sensitivity."

With the single exception of Welsh, no one has attempted to make practical use of the factor studies. Authors have been almost universally critical of the MMPI as a clinical instrument because it seemed to measure only two variables and took 12 or more scales to do the job. To a very large extent this criticism is a justifiable one although it should be noted that a number of the scales have only moderate communalities in the tables of factor loadings. An example of this is found in the recent study of Kassebaum et al.

(1959) in which 6 of the 12 scales have communalities below .50 after the extraction of three factors. These are the *L*, *F*, *Hs*, *Hy*, *Mf*, and *Pa* scales. Thus many of the scales have extremely large error variances or measure something which is unique to the particular scale. The positive results of a large number of predictive studies using these scales would support the latter interpretation. The end result of the factor studies on the one hand and the predictive studies on the other leaves the practising clinician in a state of confusion.

The present study attempts to identify further factors in samples of female NP patients at a VA hospital. One of the flaws in previous studies is that few have been replicated using the same matrix of scales. The authors have hesitated to accept more than the first two factors as significant. Some have not used the validity scales (*L*, *F*, *K*), and others have employed a variety of additional scales beyond the 9 original scales with little congruence of scales from one study to another. This study utilizes 17 scales in two samples of NP females.

METHOD

Subjects. Two samples ($N_s = 62, 85$) of female patients were used. Aside from the fact that the first group of records was collected earlier than the second, the two groups did not appear to differ. The mean age of the total group was 35.4 with an *SD* of 8.3. Length of hospitalization ranged from a few days to 10 years. The diagnostic groups represented are presented in Table 1. Mean scores on the MMPI scales for the combined sample are presented in Table 2. Tests of significance between the groups on each of the 17 scales showed a difference only on the *Mf* scale and this was small in absolute magnitude.

Scales. Seventeen MMPI scales were used in each sample. These were the 3 validating scales, the 9 clinical scales, the Taylor *A* scale (*A_T*) (Taylor, 1953), the *A* and *R* scales (*A_w* and *R_w*) (Welsh,

TABLE 1
PRIMARY DIAGNOSES OF TWO SAMPLES
OF NP FEMALES

	Sample A (N = 62)	Sample B (N = 85)
Schizophrenia	32	45
Manic-Depressive	2	6
Psychotic Depression	1	1
Schizophrenic, Post-lobotomy	3	0
Character and Personality Disorders	3	7
Neuroses	17	21
Organics	4	3
Unclassified	0	2
	62	85

1956), the Dependency scale (*Dp*) (Navran, 1954), and the Ego Strength scale (*Es*) (Barron, 1953).

Statistical technique. Raw scores were used for obtaining product-moment correlations. The factor analysis was carried out by Thurstone's centroid

method. All rotations were done by simple graphical procedures.

RESULTS

Correlation matrices for the separate samples are not presented since the only purpose in using two analyses was to assess the stability of factor loadings.

Four factors were extracted from the correlation matrix of each sample. Further factors were not extracted since the average residual value in the fourth residual matrix was .02. The largest residual value found at this time was .08. The centroid loadings are presented in Table 3. Orthogonal rotations were then carried out by simple graphical procedures. The rotated loadings are presented in Table 4. Plotting the test loadings on the factor vectors quickly indicated that a very close approximation to simple structure could be obtained by maximizing the loadings on A_w and R_w for the first and second factors;

TABLE 2
MEAN MMPI T SCORES FOR TOTAL SAMPLE
(N = 147)

L	F	K	Il _s	D	Hy	Pd	Mf	Pa	Pl	Sc	Ma
54.3	59.7	55.4	57.2	64.4	63.3	66.0	49.8	63.4	55.6	58.6	58.6

TABLE 3
CENTROID FACTOR LOADINGS FOR SAMPLES A AND B

	Factor I		Factor II		Factor III		Factor IV		R^2	
	A	B	A	B	A	B	A	B	A	B
L	-48	-46	+50	+27	-17	-32	-12	+19	52	42
F	+66	+63	+45	-40	-40	-42	+13	+19	82	77
K	-70	-66	+41	+53	+14	-14	-18	+25	71	80
Il _s	+80	+67	+34	+42	-13	-35	-34	-30	89	84
D	+69	+74	+45	+40	+36	+23	+21	+30	85	85
Hy	+60	+33	+52	+73	+05	-23	-39	-03	78	70
Pd	+65	+72	+14	-11	-17	-07	+20	+32	51	64
Mf	+26	+17	-17	+24	+16	+15	-32	-13	22	13
Pa	+70	+76	+16	-12	-17	-25	-08	+06	55	66
Pl	+91	+94	+09	-02	+27	+24	+20	+02	95	94
Sc	+89	+91	+18	-26	-17	-18	+25	+10	92	65
Ma	+42	+36	-31	-63	-42	-28	-21	-21	49	95
At	+95	+93	+02	+10	+22	+26	-03	-07	95	94
A _w	+92	+92	-14	-20	+26	+23	+17	+03	96	94
R _w	-06	+14	+66	+62	+37	+10	+11	+34	59	53
Dp	+88	+91	-22	-10	+32	+26	+18	-04	96	91
Es	-86	-80	-10	-14	-12	+18	+10	+15	77	71

TABLE 4
ROTATED FACTOR LOADINGS FOR SAMPLES A AND B

	Factor I		Factor II		Factor III		Factor IV		h^2	
	A	B	A	B	A	B	A	B	A	B
<i>L</i>	-.59	-.58	+.34	+.20	+.19	+.07	+.17	+.24	.53	.44
<i>F</i>	+.47	+.58	+.20	-.23	+.34	+.16	+.66	+.59	.81	.76
<i>K</i>	-.72	-.76	+.40	+.47	+.05	-.03	-.17	+.07	.71	.80
<i>As</i>	+.61	+.45	+.16	+.27	+.66	+.74	+.21	.00	.88	.82
<i>D</i>	+.71	+.68	+.57	+.62	+.06	+.02	+.11	+.01	.84	.85
<i>Hy</i>	+.43	+.11	+.38	+.63	+.65	+.53	+.07	-.01	.76	.69
<i>Pd</i>	+.58	+.69	+.05	+.13	+.12	-.01	+.40	+.35	.51	.62
<i>Mf</i>	+.25	+.14	-.13	+.19	+.25	+.12	-.29	+.25	.23	.13
<i>Pa</i>	+.57	+.68	+.03	-.01	+.38	+.26	+.28	+.33	.55	.64
<i>Pt</i>	+.94	+.95	+.22	+.16	+.03	+.08	+.09	-.05	.94	.94
<i>Sc</i>	+.81	+.88	+.09	-.08	+.15	+.18	+.48	+.35	.92	.94
<i>Ma</i>	+.29	+.39	-.52	-.64	+.32	+.14	+.21	+.26	.50	.65
<i>At</i>	+.94	+.92	+.10	+.22	+.25	+.16	-.01	-.14	.96	.94
<i>Aw</i>	+.98	+.97	+.02	.00	.00	.00	.00	.00	.96	.94
<i>Rw</i>	-.05	.00	+.77	+.73	-.01	.00	-.02	.00	.60	.53
<i>Dp</i>	+.97	+.94	-.04	+.04	-.06	-.09	-.07	+.05	.95	.90
<i>Es</i>	-.81	-.68	-.11	-.13	-.32	-.47	-.05	-.09	.77	.71

respectively. This procedure has the additional advantage of objectifying the process of rotation for the two samples and making the comparison of the two analyses more meaningful. Thus only the locations of the reference axes for Factors III and IV had to be determined by subjective means. This last rotation for each sample was done with Thurstone's principle of simple structure in mind. Discussion and comparison of the four factors are presented below.

Comparison of Factor Loadings

Factor I. The first factor obtained in both samples is obviously the same as that found in previous factor analyses of the MMPI. Loadings above .9 are found on *Aw*, *At*, *Dp*, and *Pt* in both samples. Correlating the pairs of loadings results in a Pearson coefficient of .98 and a rho of .96.

Factor II. Significant loadings (above .3) are found for six scales of Factor II in Sample A and for five of these same six in Sample B. In order of magnitude, with the respective loadings in parentheses, the scales are *Rw* (.77, .73), *D* (.57, .62), *Ma* (-.52, -.64), *K* (.40, .47), *Hy* (.38, .63), and *L* (.34, .20). The Pearson *r* for the 17 pairs of loadings is .87 and rho is .74. Thus considerable correspondence is found for this factor in the two

samples. The scales which have the highest loadings also indicate that the factor is highly similar to the second or *R* factor found in previous experiments.

Factor III. The third factor showed six significant loadings in Sample A but only three in Sample B. Those scales which had significant loadings in both analyses were *Hs* (.66, .74), *Hy* (.65, .53), and *Es* (-.32, -.47). The three scales which did not appear to be significant in the second sample were *Pa* (.38, .26), *F* (.34, .16), and *Ma* (.32, .14). That this difference between the two analyses is more apparent than real is indicated by a Pearson *r* of .95 between the pairs of loadings and a rho of .89. As a consequence, it was accepted that the factor was satisfactorily replicated.

Factor IV. Three scales attain significance on the fourth factor in both samples. These are *F* (.66, .59), *Sc* (.48, .35), and *Pd* (.40, .35). An additional scale, *Pa* (.28, .33), passes the significance point in Sample B. Although this factor can be accepted as similar in both samples, the correspondence is not nearly so great as in the other factors. The Pearson *r* is .70 and the rho is .58 between the sets of loadings. Fairly large differences in absolute loadings occurred on two scales, *Mf* (-.29, .24) and *K* (-.17, .07); and large relative

differences occurred on others, e.g., *Hs* from fifth highest to twelfth highest, *Pt* from ninth to fifth highest, and *Dp* from fifteenth to ninth highest.

Combined Results

Since close identity was found for the first three factors in both samples and similarity was found for the fourth, the two samples were combined and the results were reanalyzed as described previously. The correlation matrix for the combined sample, the centroid loadings for the data, and the rotated loadings are presented elsewhere.¹

Interpretation of Factors

Factor I. This factor accounts for 63.1% of the common factor variance in the table of intercorrelation. The high loadings on *A_w*, *A_T*, *Dp*, and *Pt* indicate that it is identical with factors found in previous studies. It appears to be a general maladjustment, anxiety, and/or complaint factor.

Factor II. The second factor accounts for 16.0% of the common factor variance. Since it has its greatest loading on *R_w*, we may assume that it is highly similar, if not identical, to Welsh's repressive-expressive factor. The other scales with high loadings are *D* (.62), *Hy* (.56), *Ma* (-.55), and *K* (.41). These loadings are consistent with the interpretation of a bipolar factor with repression at one extreme and expression at the other.

Factor III. This factor accounts for only 9.6% of the common factor variance in the study but showed stability from one sample to the second. In the combined sample analysis, only three scales had substantial loadings: *Hs* (.69), *Hy* (.60), and *Es* (-.38). The common variable in the *Hs* and *Hy* scales is an expression of physical symptomatology. Barron (1953) described 11 of his 68 items on the *Es* scale as relating to physical function. Thus we can tentatively label this factor as a

somatization variable. There may be other correlates of the factor such as are represented by the subtle items of the *Hy* scale (Wiener, 1948), those whose content deal with character formation rather than with physical symptomatology, but this cannot be established by the present study.

Factor IV. Factor IV accounts for 11.3% of the common factor variance in the study but is apparently stable from one sample to another. High loadings are found on five scales: *F* (.70), *Sc* (.48), *Pa* (.44), *Ma* (.37), and *Pd* (.34). These scales when found to be high on a clinical profile are usually interpreted as representing a psychotic level of functioning and/or a potential for the acting-out of impulses. The writer prefers to label the factor as an acting-out tendency at the present time since clinical experience indicates that many patients with character disorders also have high scores on these scales.

Comparison with Other Studies

Two studies were selected for comparison with the present study on the basis that similar scales were included in the matrix. Both studies (Kassebaum et al., 1959; Fisher, 1957) used male subjects, college students in the first instance and VA medical and psychiatric samples in the second. This comparison should partially answer the question as to similarity of factor structure of the MMPI between male and female subjects. The *A_w* and *R_w* factors are not discussed because they show such striking similarity in all studies.

Somatization factor. Kassebaum labeled his third factor "tender minded sensitivity," partially because of loadings on scales not included in the present study. Of the 16 scales used in both studies, the two highest loadings in both cases are on *Hs* and *Hy*, indicating that some similarity exists between the factors. A Pearson *r* of .69 was obtained for the two sets of loadings. Fisher separately analyzed a medical and a psychiatric sample, both samples being drawn from a general rather than an NP hospital. It is not surprising that his second factor resembled our somatization factor and was so labeled by him. The Pearson *r* between his medical sample and the present study was .73 and between his psychiatric sample and the present

¹ Tables of the correlation matrix, centroid loading, and rotated loadings for the combined sample have been deposited with the American Documentation Institute. Order Document No. 6465 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance \$1.25 for 35 mm. microfilm or \$1.25 for 6 × 8 in. photocopies. Make checks payable to: Chief, Photoduplication Service, Library of Congress.

study it was .83. Thus some correspondence among factors on all three studies was found. This correspondence was at a maximum when the subjects were psychiatric patients and at a minimum when the subjects were college students. Closer correspondence could probably be obtained by rotating the factors of other studies so that A_W and R_W would be at a maximum as was done in this study.

Acting-out factor. The fourth factor found in this study resembles a fifth factor found by Fisher and described as a social alienation factor with psychotic implications. The correspondence between the acting-out factor described above and Fisher's social alienation factor in the psychiatric sample is particularly close, in both cases the same five scales have loadings above .3. These are *F*, *Pd*, *Sc*, *Pa*, and *Ma*. The Pearson r for the 16 pairs of loadings is .70. As indicated previously, rotations maximizing A_W and R_W would likely result in an increased correlation.

DISCUSSION

The most important finding of this study appears to be the stability of factor loadings on replication with a similar sample of sub-jects. The somatization and acting-out factors found here each account for approximately 10% of the common factor variance, an amount which several investigators have deemed not worthy of interpretation. Nevertheless, the factors, as presently interpreted, seem to represent frequently observed behaviors in patients and appear to be worthy of objective measurement. The finding that the factors account for so little of the common factor variance is the result of the original construction of the test. For example, if scales were constructed separately to measure the different psychophysiological reactions, we might expect to find a common factor of somatization which would account for a very large amount of the common variance of the table. If, in the same table of intercorrelations, only two measures of general maladjustment were included, we would find that this factor accounted for very little of the common variance.

We may conclude that the MMPI, as presently scored with the clinical and validity scales, is overloaded with measures of general maladjustment and that other more pure

scales can be profitably constructed. The second factor found in this study can be measured fairly well by the R_W scale. The third factor (somatization) can be fairly well measured by the altitudes of the *Hs* and *Hy* scales on the MMPI profile. The loadings of these scales on Factor III are greater than their corresponding loadings on the general maladjustment factor. The five scales which have respectable loadings on Factor IV (acting-out) have high loadings on the general maladjustment factor as well. Consequently it seems that this factor cannot be so easily distinguished from general maladjustment when the clinician is faced with an individual profile. Thus if we hope to measure acting-out potential with the MMPI, it seems very desirable to construct a scale which is uncontaminated with the general maladjustment factor.

The factorial composition of the A_T scale, the *Dp* scale, and the *Es* scale shows that these measures overlap to a very considerable extent with the A_W and *Pt* scales. A substantial body of literature has grown up around several of these scales as if something unique were being measured. It would seem wise for the person who develops a new scale to do correlational studies with already existing and validated scales.

We can expect similar developments in the future, i.e., many new scales will be created and many of these will be near duplicates of those already in existence. An example of this is to be found in the recent paper by Kassebaum et al. (1959) in which 19 nonclinical scales were included in a factor analysis with the original MMPI scales. Excluding A_W and R_W , the average factor loading of the remaining 17 scales on the general maladjustment factor is .66. The same average on the second or repression factor is .31. Squaring each of these average factor loadings and summing to arrive at a communality we arrive at a figure of .53. Thus approximately 50% of the total variance of these new scales is wasted on factors which are measured much better by other scales. The supposed "nonclinical" scale is often as much affected by this contamination of general maladjustment and repression as is the clinical scale. In fact the nonclinical scale designed to arrive at the strengths of individual subjects is often a clinical scale

turned upside down. Examples drawn from the above mentioned paper include the Leadership scale with a loading of $-.85$ and the Tolerance scale with a loading of $-.80$ on the first factor. In the analysis presented here, the only extreme example is the *Dp* scale although the *Es* scale has a high loading on Factor I with a dash of Factor III (somatization) thrown in.

The above discussion can be summed up as an argument against the endless accumulation of scales on personality inventories, scales which purport to measure one thing but which actually measure something else much better. The wiser course of action would seem to be checking out each item of a new scale against certain basic scales, particularly the general maladjustment factor in any of its several forms. Each new item should at least correlate more highly with the criterion against which the scale is being developed than it does with the general maladjustment variable. Without this precautionary measure, there will be a piling up of variance associated with general maladjustment to the extent that the end result is a good measure of the wrong thing.

An argument can be made that the general maladjustment factor itself is not a pure scale and this appears to be valid. Comrey (1958) has found at least seven significant factors in an item analysis of the *Pt* scale which, as previously indicated, is a good measure of the first factor. An argument can then be made that these subcategories of general maladjustment need to be measured independently and that the general maladjustment factor is merely a global construct which would be more useful if broken down into its component parts. This matter can only be resolved by empirical investigation. A practical hindrance is the fact that only a few items are represented in each of Comrey's subscales and more would have to be found in order to make a practical evaluation of predictive power.

SUMMARY

Seventeen scales from the MMPIs of female NP subjects were factor analyzed and replicated. Four factors emerged clearly in both samples. These were labeled tentatively

as anxiety, repression, somatization, and acting-out. Comparisons were made with two factorial studies of male subjects and considerable correspondence was found for all four factors. Similarity of factor structure was greatest when NP male patients were compared with NP female patients and least when male college students were compared with female NP patients.

The results of the study seem to indicate a clear need for the construction of pure scales to measure the third (somatization) and fourth (acting-out) factors. Existing scales which relate to these dimensions of behavior are highly correlated with first and second factor scores and cannot easily be interpreted.

A further implication of the results is that careless construction of new empirical scales has resulted in near duplicates of the anxiety scales which makes them relatively useless. Additional disadvantages are that these new scales are variously named according to the particular criterion employed and that independent bodies of research tend to build up around them.

REFERENCES

- BARRON, F. An ego-strength scale which predicts response to psychotherapy. *J. consult. Psychol.*, 1953, 17, 327-333.
- COMREY, A. L. A factor analysis of items on the MMPI psychasthenia scale. *Educ. psychol. Measmt.*, 1958, 18, 293-300.
- FISHER, J. An empirical study of the relation of physical disease to body-object cathexis. Unpublished manuscript, Veterans Administration Hospital, San Francisco, California, 1957.
- KASSEBAUM, G., COUCH, A. S., & SLATER, P. E. The factorial dimensions of the MMPI. *J. consult. Psychol.*, 1959, 23, 226-236.
- NAVRAV, L. A rationally derived MMPI scale of measure dependence. *J. consult. Psychol.*, 1954, 18, 192.
- TAYLOR, JANET A. A personality scale of manifest anxiety. *J. abnorm. soc. Psychol.*, 1953, 48, 285-290.
- WELSH, G. S. Factor dimensions A and R. In G. S. Welsh & W. G. Dahlstrom (Eds.), *Basic readings on the MMPI in psychology and medicine*. Minneapolis: Univer. Minnesota Press, 1956.
- WHEELER, W. M., LITTLE, K. B., & LEHNER, G. F. J. The internal structure of the MMPI. *J. consult. Psychol.*, 1951, 15, 134-141.
- WIENER, D. N. Subtle and obvious keys for the MMPI. *J. consult. Psychol.*, 1948, 12, 164-170.

(Received January 28, 1960)

PSYCHOLOGICAL DEFICITS IN RELATION TO ACUTENESS OF BRAIN DYSFUNCTION¹

KATHLEEN B. FITZHUGH, LOREN C. FITZHUGH

New Castle State Hospital

AND RALPH M. REITAN

Indiana University Medical Center

Yates (1954, p. 374) has voiced the criticism that many investigators seem to view brain damage as a "unitary factor." Klebanoff, Singer, and Wilensky (1954) have suggested that a major reason for lack of agreement in the results of studies of psychological impairment related to organic brain damage or disease in large part may reflect differences in type, locus, or severity of the brain lesions represented in the samples studied. Birch and Diller (1959) point out that "a clear view of the evidence is made difficult, or even impossible, by the fact that the various parameters of cerebral dysfunction have not been examined systematically" (p. 188). Macroscopic and microscopic studies as reported in neurology textbooks such as Wechsler (1958) indicate that the type or severity of brain lesions may cause marked differences in the organic condition of the brain. The detrimental effects upon adaptive abilities due to acutely destructive lesions such as intrinsic tumors or cerebral vascular accidents may be more dramatic than the effects of relatively static conditions such as healed head wounds or slowly progressive conditions. The present study was designed to investigate psychological deficits in relation to acuteness of organic brain lesions.

METHOD

Subjects

Four groups, each consisting of 16 hospitalized patients, were studied. Corresponding subjects were

¹ This investigation was supported in part by Research Grant B-1468 from the National Institute of Neurological Diseases and Blindness, United States Public Health Service.

The writers are indebted to Maryellen Means for assistance with the statistical computations.

individually matched as closely as possible according to chronological age, sex, race, and years of education. Three groups were composed of patients diagnosed as having organic brain damage or disease. Diagnoses were based upon detailed medical history, electroencephalography, neurological examination, and, when further clarification was needed, angiography, pneumography, and repeated neurological examinations. The fourth, or Control group, was composed of patients in whom organic brain damage was confidently ruled out on the basis of similar, although generally less extensive, clinical diagnostic procedures.

One brain damaged group (Acute) was composed of patients who had acute neurological illnesses and whose neurological signs and symptoms were present at the time of psychological testing. These patients had experienced a specific, temporally defined, episode during which their current neurological findings had arisen, or had developed a rapidly progressive brain disease with steady progression of neurological signs. A second brain damaged group (Relatively Static) was composed of patients who had either recovered from acute neurological signs if there had been an acute onset of symptoms, or who had slowly progressive brain disease without evidence of acute or sudden onset. Among this group, the patients with sudden onset of brain dysfunction (e.g., penetrating head injury) had with the passage of time recovered from acute neurological deficits, suggesting reorganization of brain functions and a relatively static condition of the brain. The third brain damaged group (Chronic-Static) was composed of patients with chronic, long-standing brain dysfunction who were institutionalized in a state hospital for patients with neurological disorders. The diagnoses of all patients in this group included some form of epilepsy. None of the other groups included institutionalized patients. Diagnoses of the patients in the four groups are presented in Table 1.

Differences between the mean ages and mean number of years of education among the groups did not approach statistical significance. The mean ages, in years, were: Acute, 32.62 (*SD* 10.13); Relatively Static, 33.88 (*SD* 10.39); Chronic-Static, 32.88 (*SD* 10.84); and Controls, 32.38 (*SD* 10.82). Mean years of education for the groups, in the same order, were:

TABLE 1
DIAGNOSTIC DISTRIBUTIONS WITHIN BRAIN DAMAGED AND CONTROL GROUPS

Acute (<i>N</i> = 16)		Relatively Static (<i>N</i> = 16)	
Acute subdural hematoma	1	Cerebral arteriosclerosis	1
Astrocytoma	3	Chronic subdural hematoma	1
Cerebral vascular accident	3	Closed head injury	1
Glioblastoma multiforme	3	Cortical atrophy	2
Metastatic carcinoma, cortical	2	Healed cortical abscess	1
Postoperative arteriovascular malformation	1	Healed penetrating head wound	1
Preoperative meningosarcoma	1	Multiple sclerosis	5
Recent penetrating head injury	2	Posttraumatic concussion	1
		Psychomotor epilepsy	3
Chronic-Static (<i>N</i> = 16)		Controls (<i>N</i> = 16)	
Convulsive disorder due to infectious disease	3	Cancer of nasopharynx	1
Convulsive disorder (grand mal) due to unknown cause	7	Character disorder	2
Posttraumatic convulsive disorder	4	Facial laceration	1
Psychomotor epilepsy	2	Neurological complaints without CNS disease	2
		No clinical disorder found	1
		Non-CNS surgery	2
		Paraplegia	2
		Psychoneurosis	2
		Recurrent lumbar disc disorder	1
		Schizophrenic reaction	1
		Superficial occipital osteoma	1

9.69 (*SD* 2.36); 9.31 (*SD* 2.08); 9.00 (*SD* 2.83); and 9.06 (*SD* 3.03).

Procedure

All patients were administered the Wechsler-Bellevue Intelligence Scale, Form I, and seven of the measures described by Halstead (1947) as indicators of biological intelligence. The seven Halstead indicators used were those found by Reitan (1955b) to be the most sensitive for differentiating between subjects with and subjects without evidence of organic brain damage. Additionally, a composite score (Impairment Index) was computed for each subject

based upon the number of Halstead variables on which the subject's performance ranked within the range characteristic of brain damaged individuals.

In order to facilitate group comparisons and equalize variability on the several measures on each variable, the raw scores from all groups were pooled and ranked poorest to best performance. These ranks were converted to normalized standard scores (*T* scores). Since the groups had been equated by matching individuals, any two groups could be compared by calculating the mean of the *T* score differences between the corresponding individuals in the two groups. This mean difference, in turn, was evaluated by Student's *t*. Also, in order to present

TABLE 2
WECHSLER-BELLEVUE SUMMARY SCORES AND STANDARD DEVIATIONS
ACCORDING TO ACUTENESS OF LESIONS

IQ	Acute		Relatively Static		Chronic-Static		Control	
	Mean	<i>SD</i>	Mean	<i>SD</i>	Mean	<i>SD</i>	Mean	<i>SD</i>
Full	80.38	13.99	92.81	17.07				
Verbal	80.31	18.45	95.12	15.70	90.38	14.68	108.81	10.00
Performance	84.44	13.88	91.81	18.27	88.88	15.13	105.38	10.00
					93.75	13.05	111.38	11.15

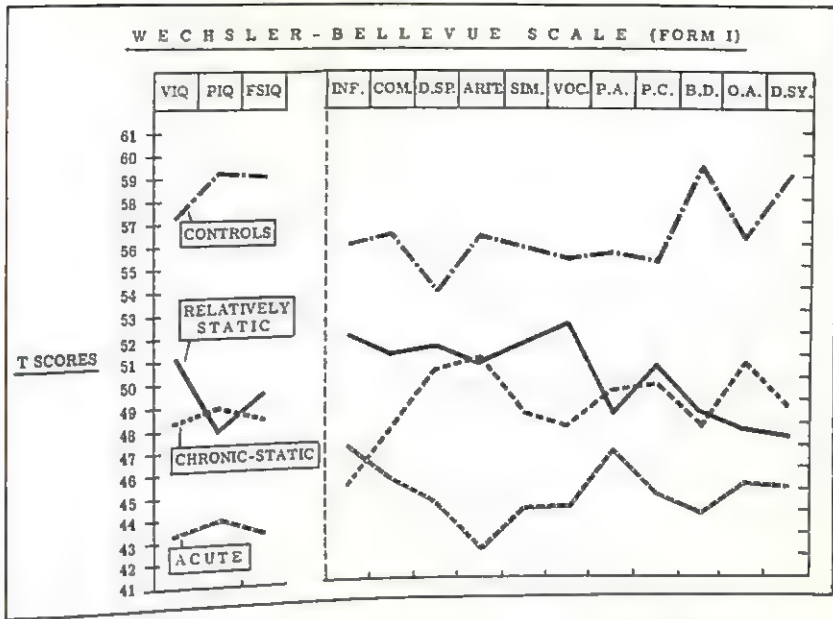


FIG. 1. Graphic presentation of mean *T* score values on Wechsler-Bellevue variables for control group and three brain damaged groups.

the intelligence quotients in a familiar manner, mean scores and standard deviations for these variables were computed from the raw scores.

RESULTS

Means and standard deviations of the Full, Verbal, and Performance scale IQ scores are presented in Table 2. Only the Acute group's mean IQ scores were consistently below the range of 90-109.

The general trends of performances of the four groups on the Wechsler-Bellevue variables may be seen in Figure 1. The Control group performed at levels consistently superior to those of the brain damaged groups.

The mean IQ score differences between Controls and brain damaged groups were significant at, or well beyond, the .05 level (see Table 3). Among the brain damaged groups, mean scores for the Acute lesion group were generally inferior to those of the two static lesion groups. Mean Verbal IQ of the Relatively Static group was higher than that of the Acute group ($p < .05$), and mean Performance IQ of the Chronic-Static group was higher than that of the Acute group ($p < .05$). Two of the three brain damaged groups (Acute and Chronic-Static) obtained slightly higher Performance than Verbal mean IQ scores (see Table 3).

On only 2 of the 11 Wechsler subtests did the mean difference scores between Controls and Acutes fail to exceed the .01 level of significance; and the mean difference scores were significant beyond the .05 level on those two subtests (Digit Span and Picture Arrangement). Comparisons between Controls and each of the static lesion groups also yielded significant differences on several subtests (see Table 3).

The general trends of the four groups on the Halstead Neuropsychological measures may be seen in Figure 2. As with the Wechsler-Bellevue, the Control group performed at levels consistently exceeding those of the brain damaged groups. Among the latter groups, the two static lesion groups performed at fairly comparable levels, although their mean scores generally exceeded those of the Acute group.

On all but two of the eight Halstead variables, mean difference scores were significant beyond the .001 level when the Control and Acute groups were compared. On the two remaining variables, Speech-Sounds Perception and Finger Oscillation, the Control and Acute groups were differentiated beyond the .01 and .05 levels, respectively (see Table 4).

Controls were differentiated from Chronic-

TABLE 3
t RATIOS BASED UPON DIFFERENCES BETWEEN EQUATED
PAIRS ON WECHSLER-BELLEVUE VARIABLES

Test Variable	Control vs. Acute	Control vs. Relatively Static	Control vs. Chronic	Acute vs. Relatively Static	Chronic vs. Acute	Chronic vs. Relatively Static
Full IQ	5.14****	3.41***	4.22****	1.87	1.58	.43
Verbal IQ	4.20****	2.32*	3.17***	2.37*	1.31	1.15
Performance IQ	6.13****	3.78***	4.47****	1.16	2.21*	.39
Information	3.55***	1.36	4.34****	1.16	.50	3.04***
Comprehension	3.66***	1.70	2.95***	1.55	.66	1.05
Digit Span	2.20*	1.06	1.39	1.89	1.69	.29
Arithmetic	4.75****	1.81	1.88	2.91**	2.19	.04
Similarities	4.28****	1.73	2.47*	2.33*	.96	.99
Vocabulary	4.10****	.92	2.53*	2.00	.93	1.85
Picture Arrangement	2.86**	2.47*	2.10	.44	1.03	.33
Picture Completion	4.67****	1.17	1.94	1.57	1.55	.26
Block Design	6.38****	3.71***	5.18****	1.97	1.54	.29
Object Assembly	4.36****	3.52***	2.14*	.60	2.03	1.16
Digit Symbol	5.16****	3.92***	5.17****	.65	1.22	.38

* $p < .05$.
** $p < .02$.
*** $p < .01$.
**** $p < .001$.

Statics on all Halstead variables except Speech-Sounds Perception; and Controls were significantly differentiated from Relatively Statics on five of the eight Halstead variables. Every brain damaged group was differentiated from Controls on the composite measure, Im-

pairment Index, at levels exceeding the .01 level. Such differentiation is consistent with the findings of Reitan (1959a) on a heterogeneous group of brain damaged patients.

On the 22 test variables studied the two static groups differed significantly from each other on only one, the Information subtest of the Wechsler-Bellevue. This particular difference may be considered suggestive of the effects of institutionalization upon the Chronic-Static group. In contrast, the Acute group performed significantly less well than one or both of the static groups on several variables. Differentiation occurred at levels exceeding the .05 level on the Wechsler-Bellevue variables of Arithmetic, Similarities, Verbal IQ, and Performance IQ. The Halstead Indicators of Memory and Location variables of the Tactual Performance Test, and the Seashore Rhythm Test differentiated Acutes from one or both of the static groups at levels exceeding the .05 level of significance.

DISCUSSION

As Rosvold (1959) pointed out recently: studies with respect to the effect of brain damage on general intelligence, though more rigorous than in the past, are no more in agreement than were ear-

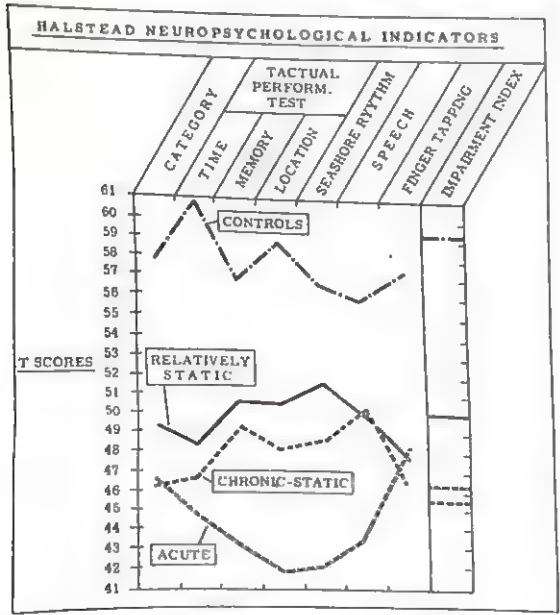


FIG. 2. Graphic presentation of mean T score values on Halstead Neuropsychological measures for control group and three brain damaged groups.

TABLE 4

RATIOS BASED UPON DIFFERENCES BETWEEN EQUATED PAIRS ON
HALSTEAD NEUROPSYCHOLOGICAL INDICATORS

Indicator	Control vs. Acute	Control vs. Relatively Static	Control vs. Chronic	Acute vs. Relatively Static	Chronic vs. Acute	Chronic vs. Relatively Static
Category	4.90****	2.61**	3.90***	.82	.04	.87
TPT Time	6.02****	5.51****	5.62****	.92	.72	.57
TPT Memory	5.20****	1.85	2.45*	2.20*	1.64	.62
TPT Location	5.72****	3.04***	3.96***	2.72**	2.15*	.86
Rhythm	4.66****	1.94	2.59*	3.41***	2.10	1.26
Speech	3.38***	1.47	1.46	1.53	1.58	.11
Finger Oscillation	2.55*	3.27***	4.50****	.04	.42	.54
Impairment Index	6.50****	3.31***	5.17****	1.27	.25	1.83

* $p < .05$.** $p < .02$.*** $p < .01$.**** $p < .001$.

lier studies, some of which claimed deterioration, others not (p. 434).

The basis for disagreement may relate in part to differences in the types of brain damage studied. Many studies in this area have used subjects with brain damage resulting from head injuries (Aita, Armitage, Reitan, & Rabinovitz, 1947; Milner, 1956; Ross, 1958; Teuber & Mishkin, 1954; Weinstein & Teuber, 1957). (These and other references in this section are illustrative rather than exhaustive.) In some instances the patients have been in the intermediate recovery period when tested (Aita et al., 1947), in others (notably Teuber and his co-workers) at least several years have elapsed since the head trauma occurred, and in temporal lobe epilepsy (Milner, 1956) brain damage probably occurred in most instances at birth or in childhood. In terms of EEG tracings, evidence has been presented to indicate that many patients recover from head injuries, gradually approaching more normal results (Jasper, Kershman, & Elvidge, 1945).

The variation in type and severity of brain damage associated with developing brain disease processes has been well documented (Wechsler, 1958). Relatively fewer investigators have used such patients for psychological evaluations (Battersby, Krieger, & Bender, 1955; Halstead, 1947; Morrow & Mark, 1955; Reitan, 1955a). Again, the diversity

of these conditions, either within or between diagnostic categories, is well known to neurologists, neurological surgeons, and neuropathologists. It is not beyond the scope of reasonable possibility that different diagnostic conditions may reflect themselves differently in psychological testing. In fact, Reitan (1959b) has recently demonstrated that inferences based on psychological test results alone (without reference to anamnestic material or other findings) identify patients with head trauma, brain tumors, cerebrovascular accidents, and other diagnostic conditions at levels far exceeding chance expectancy. The dependent variables, or psychological tests, have also frequently varied from one study to another among different investigators. This factor also would contribute a certain amount of variance to the conclusions drawn.

Because of the impossibility of simultaneous manipulation of the many factors that are probably relevant, the results of any single study in this area must be viewed as tentative. The present findings, however, agree with certain others in which the same instruments were used in indicating that the effects of brain damage may be measured reliably (Fitzhugh, Fitzhugh, & Reitan, in press; Kløve, 1959; Kløve & Reitan, 1958; Reitan, 1955a, 1955b, 1958). Additionally, among the brain damaged groups consistent trends were observed revealing greater psychological

impairment in patients who suffered from acute organic brain damage or disease than in groups with relatively static organic damage. The results suggest that the nature of the brain lesion at the time of psychological testing is an important variable and should be considered in studies of psychological deficits in association with brain damage.

SUMMARY

One Control group and three brain damaged groups, each composed of 16 patients, were compared on the Wechsler-Bellevue Intelligence Scale variables and eight Halstead Neuropsychological indicators in order to investigate psychological impairment in relation to acuteness of organic brain dysfunction. The Control group's performances consistently exceeded the performances of the brain damaged groups. Also, two static lesion groups (one institutionalized) rather consistently performed at levels superior to the levels of the Acute lesion group. The results suggested that acuteness of the organic brain lesions is an important variable to be considered in studies of psychological deficits among brain damaged subjects.

REFERENCES

- ALTA, J. A., ARMITAGE, S. G., REITAN, R. M., & RABINOVITZ, A. The use of certain psychological tests in the evaluation of brain injury. *J. gen. Psychol.*, 1947, 37, 25-44.
- BATTERSBY, W. S., KRIEGER, H. P., & BENDER, M. B. Visual and tactile discriminative learning in patients with cerebral tumors. *Amer. J. Psychol.*, 1955, 68, 562-574.
- BIRCH, H. G., & DILLER, L. Rorschach signs of "organicity": A physiological basis for perceptual disturbances. *J. proj. Tech.*, 1959, 23, 184-197.
- FITZHUGH, L. C., FITZHUGH, K. B., & REITAN, R. M. Performance of hospitalized alcoholic subjects on measures of adaptive abilities and intellectual functioning. *Quart. J. Stud. Alcohol*, in press.
- HALSTEAD, W. C. *Brain and intelligence: A quantitative study of the frontal lobes*. Chicago: University of Chicago Press, 1947.
- JASPER, H., KERSHMAN, J., & ELVIDGE, A. Electroencephalography in head injury. *Res. Publ. Ass. Res. Nerv. Ment. Dis.*, 1945, 24, 388-420.
- KLEBANOFF, S. C., SINGER, J. L., & WILENSKY, H. Psychological consequences of brain lesions and ablations. *Psychol. Bull.*, 1954, 51, 1-41.
- KLØVE, H. Relationship of differential electroencephalographic patterns to distribution of Wechsler-Bellevue scores. *Neurology*, 1959, 9, 871-876.
- KLØVE, H., & REITAN, R. M. Effect of dysphasia and spatial distortion on Wechsler-Bellevue results. *AMA Arch. Neurol. Psychiat.*, 1958, 80, 708-713.
- MILNER, BRENDA. Psychological defects produced by temporal lobe excision. *Res. Publ. Ass. Res. Nerv. Ment. Dis.*, 1956, 36, 244-257.
- MORROW, R. S., & MARK, J. C. Intelligence of patients autopsied for brain damage. *J. consult. Psychol.*, 1955, 19, 283-289.
- REITAN, R. M. Certain differential effects of left and right cerebral lesions in human adults. *J. comp. physiol. Psychol.*, 1955, 48, 474-477. (a)
- REITAN, R. M. Investigation of the validity of Halstead's measures of biological intelligence. *AMA Arch. Neurol. Psychiat.*, 1955, 73, 28-35. (b)
- REITAN, R. M. Validity of the Trail Making Test as an indicator of organic brain damage. *Percept. mot. Skills*, 1958, 8, 271-276.
- REITAN, R. M. The comparative effects of brain damage on the Halstead impairment index and the Wechsler-Bellevue scale. *J. clin. Psychol.*, 1959, 15, 281-285. (a)
- REITAN, R. M. *The effects of brain lesions on adaptive abilities in human beings*. Indianapolis: Author, 1959. (Mimeo) (b)
- ROSS, A. O. Brain injury and intellectual performance. *J. consult. Psychol.*, 1958, 22, 151-152.
- ROSVOLD, H. E. Physiological psychology. *Ann. Rev. Psychol.*, 1959, 10, 415-454.
- TEUBER, H. L., & MISHKIN, M. Judgment of visual and postural vertical after brain injury. *J. Psychol.*, 1954, 38, 161-175.
- WECHSLER, I. S. *A textbook of clinical neurology*. (8th ed.) Philadelphia: Saunders, 1958.
- WEINSTEIN, S., & TEUBER, H. L. Effects of penetrating brain injury on intelligence test scores. *Science*, 1957, 125, 1036-1037.
- YATES, A. J. The validity of some psychological tests of brain damage. *Psychol. Bull.*, 1954, 51, 359-379.

(Received February 1, 1960)

TEMPORAL AND EMOTIONAL FACTORS IN THE SELECTIVE RECALL OF DREAMS

ROSALEA ANN SCHONBAR

Teachers College, Columbia University

There is increasing evidence that everyone dreams approximately five times every night (Aserinsky & Kleitman, 1955; Dement, 1955; Dement & Kleitman, 1957; Dement & Wolpert, 1958). Yet even when there is motivation to recall dreams, as in psychotherapy or research, some people recall no dreams at all (Schonbar, 1959), and none report anywhere near the maximum possible. Why are so many dreams lost? What characterizes those which are remembered?

The present study is based upon the observations of the investigators cited above that dreams occur intermittently during the whole sleep cycle (except in the first hour), that they are associated with lighter phases of sleep as indicated by EEG, and that they tend to get longer during the night. The study grew out of Freud's theories concerning the function of dreams and of a theory arising from Freud's.

According to Freud (1949, 1953, 1957), a major function of the dream is to preserve the sleep of the dreamer. In sleep, the ego gives up its cathexes in both the external and internal worlds; the unconscious or id, however, does not sleep, and, because of the relaxation of the censorship of the somnolent ego, becomes more able to intrude its desires upon the individual. Were these wishes to be expressed in undisguised form, they would create sufficient anxiety to awaken the sleeper. Freud therefore considers the dream to be an economical compromise, with the forbidden impulses allowed expression in disguised form, experienced as objective rather than subjective events, thus not demanding full censorship, but allowing sleep to continue.

If the demand made by the unconscious is too great, so that the sleeping ego is not in a position to ward

it off by the means at its disposal, it abandons the wish to sleep and returns to waking life . . . every dream is an *attempt* to put aside a disturbance of sleep . . . This attempt can be more or less completely successful; it can also fail—in which case the sleeper wakes up, apparently aroused by the dream itself (Freud, 1949, pp. 56-57). (Quoted by permission of Norton)

Such failures are identified as anxiety-dreams (Freud, 1953).

Related to a part of Freudian theory is Gutheil's view (1951) that the dream serves to protect the integrity of the ego. Gutheil proposes that dreams are most likely to occur just as the individual is falling asleep and just as he is awakening, so that the ego is able to make gradual adjustments to the differing demands of the two states and is not pressed into abrupt, and possibly disintegrating, changes in function. Gutheil predicts further that dreams from the falling asleep period are less likely to be remembered than those from just before waking because of the long period of unconsciousness which intervenes in the former case.

For the most part, the above views grew out of the analysis of retrospectively recalled dreams, mostly of patients in psychotherapy or of Freud himself. There was no way of knowing then that these dreams were merely a sample of a much larger and determinable number. The present study is concerned with testing some propositions based in theory, but in terms of selective recall, since this is the significant factor in what is available to us under nonlaboratory conditions.

The first two hypotheses to be tested are that more dreams are remembered as having preceded a waking period than as having preceded continued sleep, and that proportionately more dreams are remembered as oc-

TABLE 1
NUMBER OF DREAMS IN EACH TEMPORAL AND FEELING CATEGORY
FOR GROUPS H ($N = 19$) AND L ($N = 19$)

Time	Group	Feelings				Total	
		Neutral	Pleasant	Unpleasant			
				Nonanxious	Anxious		
Indeterminate ^a	H	47	18	10	22	97	115
	L	7	5	3	3	18	
Awoke Dreamer ^b	H	14	4	14	21	53	65
	L	6	1	1	4	12	
Awoke Dreamer in Morning	H	1	1	1	3	6	8
	L	0	0	0	2	2	
Morning	H	21	8	14	14	57	73
	L	12	2	0	2	16	
Total	H	83	31	39	60	213	261
	L	25	8	4	11	48	
	Combined	108	39	43	71	261	

^a Includes 7 dreams identified as F (Falling asleep).

^b Includes 4 dreams identified as FW (Awoke the dreamer while he was falling asleep).

Hypothesis 4. Of the dreams having feelings other than N ascribed to them, the number of P and U feelings was tested against a 50-50 expectancy. For Group H, χ^2 was 37.7, $p < .0001$. For Group L, χ^2 was 2.14, $p > .05-.10$. The recalled dreams of frequent recallers are not only characterized by having more feelings than neutral emotional components, but also by more unpleasant than pleasant feelings. The reported dreams of infrequent recallers, on the other hand, are not only more emotionally neutral, but, when feelings are remembered, they are not more likely to be unpleasant than pleasant.

Hypothesis 5. Dreams which awakened the Ss (W, MW) were divided into those with anxiety and those without; for Group H, this distribution was tested against the distribution of anxiety and its absence in all other dreams. χ^2 was 65.26, significant beyond .0001. This hypothesis was not tested for Group L because the frequencies were too small, although they fell in the predicted direction. It may be concluded that, while anxiety is not experienced in a majority of Group H's W and MW dreams, a significantly greater proportion of anxious feelings is associated

with these dreams than with those which did not awaken the individual.

Hypothesis 6. The prediction that dreams recalled as having occurred at indeterminate times during the night and followed by continuous sleep contained more pleasant feelings than other dreams was tested in the same manner for Group H, Group L not having high enough frequencies for meaningful testing. χ^2 was 7.15, significant at the .003 level. Again, although most of these dreams were not remembered as pleasant, they were remembered as being significantly more pleasant than all others.

Hypothesis 7. Group H designated 83 of its 213 dreams as being neutral in feelings; Group L, 25 of its 48. When the distribution of N and other feelings in Group L was compared with an expected frequency based upon the distribution in Group H, χ^2 was 5.28, significant at the .01 level. In addition, this hypothesis received inadvertent support in the testing of Hypothesis 3. People who recall fewer dreams remember proportionately more of them as neutral in feeling than do people whose dream recall is greater.

DISCUSSION

The experimental evaluation of any theory is a two-step process. First, it must be determined whether the events or processes assumed, implied, or predicted by the theory are confirmed by observation. Second, and more difficult, is the matter of whether the given theory explains the observations more adequately or more economically than other alternatives.

For example, Freud said, "*Dreams are the GUARDIANS of sleep . . .*" (1953, p. 233). They occur, then, when sleep is endangered. And investigations into the physiological correlates of dreaming have demonstrated that dreaming does occur during periods of lighter sleep. A necessary factual prerequisite has thus been established. But it does not necessarily follow that these "new laboratory experiments . . . have corroborated Freud's brilliant guess" (Robinson, 1959, p. 52). Freud would maintain that unconscious wishes have brought about the lighter sleep by striving for expression, that the dream puts a stop to this so that sleep may continue. But it may also be, of course, that the dream itself somehow interferes with the depth of sleep. Thus, the discovery of the correlation of dreaming and lighter phases of sleep is a necessary but not sufficient condition for support of Freud's theory.

The research reported here is similarly concerned with establishing observationally the verification of some assumptions or implications of Freud's views. For example, it was found that, in general, dreams were not better remembered simply because they precede a waking period, and that, for those who recall relatively more dreams, the period just before normal waking did not produce more than its proportionate share of recalled dreams. Not only do these findings fail to support Gutheil's statements, but they cast considerable doubt upon any notion that dream recall is primarily dependent upon factors similar to those studied by Ebbinghaus and others: recency, opportunity for recitation, greater length, possibly greater rationality, and lack of opportunity for retroactive inhibition. Seemingly, more dynamic selective factors are operating.

Similarly, if dreams merely repeat events of the day before, or are arbitrary representations of digestive processes, or responses to fortuitous external stimulation, then it should not have been found, as it was for Group H, that the dreams were accompanied by emotion more often than not, or that the emotion was more often unpleasant than pleasant. But these findings are necessary to a theory which postulates that the dream represents conflict which is important to the dreamer.

The finding that dreams which awakened the sleeper were proportionately more often identified as anxiety dreams than were dreams followed by further sleep or normal waking directly supports one of Freud's theoretical statements. The finding that dreams followed by continued sleep contain more pleasant feelings than do other dreams is somewhat ambiguous. It would seem that these dreams might be the most "successful" in Freud's sense—at least of remembered dreams—not disturbing sleep, and possibly most disguised in the sense of being remembered as enjoyable; one cannot, however, wholly discount the likely possibility that the unpleasant aspects of these dreams became the victim of further repression during sleep, but there is no way of finding out from these data. It should be emphasized that even these dreams are not characterized as pleasant; it is rather that pleasant feelings are more likely to be associated with them than with others.

This study has replicated, in a nonclinical situation and with a nonpatient sample, the clinical procedure in which people report dreams which they remember, thus providing material similar to that upon which Freud and other psychoanalysts have made their observations. The findings of this study, at least with people who tend to recall dreams, confirm the validity of the observations upon which some aspects of Freud's dream theory were built.

There was only one prediction concerning the relationship between feelings in dreams and the greater or lesser tendency to recall dreams. This was that recallers of relatively few dreams would also remember them as being more neutral in feeling than would more frequent recallers, and this was confirmed. In addition, dreams recalled by the

low recallers occurred disproportionately more often from the period preceding morning waking and, if feelings were attributed to the dreams, they were not more likely to be unpleasant than pleasant. It may be that a larger sample of dreams from infrequent recallers might have reversed the latter finding but, as it stands, it would seem that people who recall few dreams also recall them as being fairly bland or less unpleasant than do people who recall more frequently. It is possible that the less frequent recallers do not have so many conflicts so that their dreams are, in fact, more bland. But it is at least equally possible, and seems more likely, that people who repress more of their available experience, as in forgetting dreams, reveal this repression rather generally by also toning down the affect. Previous research (Schonbar, 1959; Singer & Schonbar, 1959) has found that dream recall is positively related to manifest anxiety, but, since the latter was measured by conscious self-report, the dilemma is merely emphasized rather than resolved. A similar question arises concerning the finding (Singer & Schonbar, 1959) that repression (MMPI R scale) and dream recall are negatively correlated. But there is also evidence (Schonbar, 1959) that people who recall no dreams also tend not to recall even the process of dreaming. It would thus seem that these people exhibit a pattern of repression or lack of awareness of the presence and nature of their own dream processes. A pattern is suggested, wherein people who tend to be aware of their own internal experience remember more dreams and more of the affect associated with them, while less aware individuals remember fewer dreams and blander affect.

In summary, then, the findings of this study support the underpinnings of some aspects of Freud's theory of dreams and fail to support Gutheil's contention. From the more difficult point of view of theoretical adequacy, it is worth noting that, while Freud attributed the memory of anxious (and possibly of unpleasant) dreams and the existence of dreams which disturb sleep to a breakdown or failure of ego function, other analytic theorists (Fromm, 1951; Hadfield, 1954) would give credit for these events to a successful break-

through of self-realizing, insight-producing forces. But the same kind of substructure of intrapsychic conflict is assumed by them as by Freud, and the findings of this research, therefore, also offer confirmation of their views.

SUMMARY

Forty-five graduate students in education turned in reports on recalled dreams every day for 4 weeks. On these reports were included information concerning the time during the sleep cycle when the dream occurred, and what kinds of feelings were associated with it. The total group was divided into two, above and below the median in dream recall. One-tailed chi square tests were used to test predictions based primarily upon formulations drawn from Freud's theory of dreams. It was found for both groups that dreams preceding a waking period are not better remembered than dreams followed by continued sleep, that dreams which awaken the sleeper are proportionately more often associated with anxiety than dreams which do not, and that dreams which are followed by continued sleep are recalled as proportionately more pleasant than dreams followed by any kind of waking. For the frequent recallers, it was also found that dreams are more often remembered as having had emotional components than as having been neutral, that the feelings are more often unpleasant than pleasant, and that the period just before normal morning waking does not produce more than its proportionate share of remembered dreams. For the group which was low in recall, the recalled dreams did not contain more emotional than neutral attributes, and feelings were not more unpleasant than pleasant; more dreams were remembered by this group from the period just preceding normal waking than would be expected. In addition, a direct comparison of the two groups revealed, as predicted, that the low recall group had significantly more neutral dreams than the high recall group. In general, it was concluded that the findings of this study support some of the propositions in Freud's theory of dreams. The study is not seen as a crucial test of theory.

REFERENCES

- ASERINSKY, E., & KLEITMAN, N. Two types of ocular motility occurring in sleep. *J. appl. Physiol.*, 1955, 8, 1-10.
- DEMENT, W. Dream recall and eye movements during sleep in schizophrenics and normals. *J. nerv. ment. Dis.*, 1955, 122, 263-269.
- DEMENT, W., & KLEITMAN, N. The relation of eye movements during sleep to dream activity: An objective method for the study of dreaming. *J. exp. Psychol.*, 1957, 53, 339-348.
- DEMENT, W., & WOLFERT, E. A. The relation of eye movements, body motility, and external stimuli to dream content. *J. exp. Psychol.*, 1958, 55, 543-554.
- FREUD, S. *An outline of psychoanalysis*. (Trans. by J. Strachey) New York: Norton, 1949.
- FREUD, S. The interpretation of dreams. In J. Strachey (Ed.), *The standard edition of the complete psychological works of Sigmund Freud*. Vols. 4 & 5. London: Hogarth, 1953.
- FREUD, S. A metapsychological supplement to the theory of dreams. In J. Strachey (Ed.), *The standard edition of the complete psychological works of Sigmund Freud*. Vol. 14. London: Hogarth, 1957. Pp. 222-235.
- FROMM, E. *The forgotten language*. New York: Rinehart, 1951.
- GUTHRIE, E. A. *The handbook of dream analysis*. New York: Liveright, 1951.
- HADFIELD, J. A. *Dreams and nightmares*. London: Penguin, 1954.
- RAMSEY, G. V. Studies of dreaming. *Psychol. Bull.*, 1953, 50, 423-455.
- ROBINSON, L. W. What we dream—and why. *NY Times Mag.*, 1959, 47(Feb. 15), 52.
- SCHONBAR, ROSALEA A. Some manifest characteristics of recallers and nonrecallers of dreams. *J. consult. Psychol.*, 1959, 23, 414-418.
- SINGER, J. L., & SCHONBAR, ROSALEA A. Correlates of daydreaming: The dimension of self-awareness. *Amer. Psychologist*, 1959, 14, 393. (Abstract)
- SUTCLIFFE, J. P. A general method of analysis of frequency data for multiple classification designs. *Psychol. Bull.*, 1957, 54, 134-137.

(Received February 1, 1960)

ANXIETY, PREGNANCY, AND CHILDBIRTH ABNORMALITIES¹

ANTHONY DAVIDS, SPENCER DEVAULT, AND MAX TALMADGE

Brown University and Emma Pendleton Bradley Hospital

Since the development of a rather simple instrument for assessing manifest anxiety (Taylor, 1953), there has been an epidemic of psychological studies concerned with the role of anxiety in a wide range of experimental situations. Here, we will not attempt to survey this vast literature. We wish merely to point out that these studies of anxiety have been conducted mainly in laboratory and academic research settings, and little use has been made of the instrument in clinical or "real life" situations. The developers of the instrument and many of their followers have stated (Taylor, 1956) that they are not really concerned with measuring anxiety, but are interested in obtaining a measure of "drive." This concept of drive is viewed within the framework of Hullian learning theory. According to this theory, all habit tendencies activated by a given stimulus are considered to be multiplied by the total drive then operating. Employing the Manifest Anxiety Scale (MAS) to provide a measure of drive strength, the performances of subjects selected on the basis of high or low anxiety scores have been compared on such measures as eyelid conditioning (Hilgard, Jones, & Kaplan, 1951; Spence & Farber, 1953; Spence & Taylor,

1951; Taylor, 1951), verbal learning (Lucas, 1952; Montague, 1953; Taylor & Spence, 1952), word association (Davids & Eriksen, 1955), and various other more complex tasks (Farber & Spence, 1953; Wesley, 1953; Westrope, 1953).

There have been some attempts to assess the clinical validity of the MAS (Buss, Wiener, Durkee, & Baer, 1955; Gleser & Ulett, 1952; Hoyt & Magoon, 1954; Kendall, 1954), and in general it does seem to be associated with clinical evaluations of anxiety. Moreover, Eriksen and Davids (1955) reported finding significant personality differences between subjects who scored high or low on the MAS, and also differences in psychological defense mechanisms. More specifically, it was found, in a group of male college students, that subjects who were high on the MAS were also pessimistic in their outlook on life and were relatively low on utilization of the mechanism of repression according to the evaluation of an experienced psychoanalyst.

It seems, then, that the MAS has demonstrated utility as a research instrument and has generated considerable interesting research. However, since most personality theorists place great emphasis on anxiety as a motivating factor in life adjustment, and since it is a well established fact that anxiety plays a crucial role in the formation of psychopathology, it seems worthwhile to conduct further research on the clinical utility of this objective instrument for assessing manifest anxiety.

At present, there appears to be increasing research interest in the effects of anxiety and stress on the psychological course of pregnancy and the influence that emotional tur-

¹This study was made possible by a research grant, B-2356, from the National Institute of Neurological Diseases and Blindness, United States Public Health Service, awarded to the Brown University Institute for Research in the Health Sciences. The present report stems from an ancillary study to the National Collaborative Project, conducted locally at the Providence Lying-In Hospital, which is investigating perinatal factors in child development. We wish to express our appreciation to Glidden Brooks, who is Director of the Research Institute at Brown University, for facilitating this study. Also, we are indebted to the clinic staff of the Providence Lying-In Hospital for their cooperation and assistance.

moil during pregnancy may have on the subsequent adjustment of the offspring. In a study of physical and mental handicaps following disturbed pregnancy, Stott (1957) suggested that prenatal influences were to blame. In studying a group of mentally defective children, he found that in a large proportion of the cases there had been marked emotional stress during pregnancy, as a result of family conflicts and personal unhappiness. In a recently reported study of the influence of prenatal maternal anxiety on emotionality in rats, Thompson (1957) tested and confirmed the hypothesis that "emotional trauma undergone by female rats during pregnancy can affect the emotional characteristics of the offspring."

The plan of the research program from which the present report derives is to use a variety of psychological procedures to study emotional factors in pregnant women. This report, however, is concerned specifically with findings obtained from the MAS administered to a group of women during pregnancy and to a group of women during pregnancy and readministered soon after delivery of their children.

METHOD

The subjects of this investigation were 48 pregnant women who were studied at the Clinic of the Providence Lying-In Hospital. They are a representative sample of a larger group of women who were studied in the course of a pilot study conducted by a team of medical and scientific investigators who were engaged in a collaborative project on perinatal factors in child development. The women were seen for individual psychological testing during the course of a routine visit to the clinic, which in most cases was at approximately the seventh month of pregnancy. Of the group of 48 women, 20 returned for a routine physical checkup at approximately 6 weeks following childbirth, while the other 28 women failed to return for this scheduled hospital visit. The 20 patients who were seen twice will be labeled Group I, and the 28 women who were seen only during pregnancy constitute Group II.

In the course of the large scale investigation, voluminous data were gathered for each patient. As part of the assessment, they were administered a comprehensive battery of psychological tests. Included in this assessment procedure was the 50-item MAS, which is the focus of the present report. In Group I, the MAS was administered both during and following pregnancy, while in Group II it was administered only during pregnancy. On the basis of the official hospital records, it was possible to classify each patient's delivery room record as "normal"

or as indicating some "abnormality or complication." In Group I, there were 13 patients in the normal category and 7 patients in the abnormal category. In Group II, the subdivisions were 12 normal deliveries and 16 with abnormalities or complications.

The patients in both groups were of "normal" intelligence. As measured by the Wechsler-Bellevue Intelligence Scale, the mean IQ in Group I was 101 and the mean IQ in Group II was 95. Moreover, in both groups the mean age was 25 years and ranged from 17 to 40 years. Thus, although no attempt was made to match the patients in the two groups, it happened that the groups were of very similar age and IQ, and in regard to these two variables it seems probable that they are representative of pregnant women who are being studied at various clinics throughout the country.

RESULTS AND DISCUSSION

Now let us consider the findings from the MAS. In Group I, on the first testing, the normal subgroup obtained a mean manifest anxiety score of 16.5, which is significantly lower ($t = 2.19$, $p = .05$) than the mean of 23.5 in the abnormal subgroup. Examination of the ranges of the manifest anxiety scores in the two subgroups further evidences the greater anxiety in the abnormal group, with their scores ranging from 14 to 37, as compared with scores ranging from 8 to 26 in the normal group. Thus, both the mean scores and the spread of the individual scores reveal the abnormal delivery group to have been relatively high on manifest anxiety according to their own avowal of feelings and symptoms during pregnancy. In analyzing the results from the second testing of the patients in Group I, it is noteworthy that the level of manifest anxiety decreased in both subgroups following pregnancy, with a mean of 15 in the normal subgroup and a mean of 18.3 in the abnormal subgroup. Although the group that experienced difficult deliveries continued to score higher on manifest anxiety than did the group who had normal delivery room experiences, the nonsignificant difference ($t = .70$) was not as pronounced as it was when the women were in a state of pregnancy.

The findings in regard to manifest anxiety in Group II were remarkably similar to those obtained in Group I. In this second group of patients, the mean MAS score in the normal subgroup was 16, which is significantly lower ($t = 2.39$, $p = .05$) than the mean score of 23.6 in the abnormal subgroup. Again, the

range of MAS scores from 4 to 30 in the normal subgroup was noticeably lower than the range from 12 to 38 in the abnormal subgroup. Thus, in both samples studied in this research, it was found that women who were later to experience complications in delivery or were to give birth to children with abnormalities tended to report a relatively high amount of disturbing anxiety while they were pregnant.

In considering these findings, it should be emphasized that at present we have no information regarding the causes or reasons underlying the higher MAS scores in the abnormal subgroup. One possibility is that the obstetricians may have anticipated abnormalities or complications, and may have conveyed this information to the patients. However, this possibility does not seem too likely, as for the majority of these patients the psychological assessment was conducted during their first visit to the clinic. That is, these women did not have private obstetricians who followed their medical progress throughout the pregnancy, but were being seen for their first medical examination at a rather late stage of their pregnancy. Future examination of sociological, medical, and past history data on these clinic patients may provide some understanding of causative factors, and greater understanding in this regard may well come from comparisons of clinic and private patients. One other point that should be made at this time, however, is that there was no difference between the two subgroups in regard to the number of patients for whom this was the first delivery. The mean number of previous pregnancies and previous deliveries was practically identical in the normal and abnormal subgroups.

It is also interesting to note that the mean MAS scores of about 16, obtained in the normal subgroups both during and after pregnancy, are very similar to the mean MAS scores obtained previously in relatively large samples of female college undergraduates (Smith, Powell, & Ross, 1955; Taylor, 1953). The present findings suggest, therefore, that, as a group, pregnant women who will later experience normal childbirth do not differ from normal nonpregnant college females in the avowal of manifest anxiety, but pregnant

women who are likely to experience childbirth abnormalities later are significantly higher on manifest anxiety than are other groups of pregnant and nonpregnant women.

The results of this preliminary study, which should be regarded as tentative and in need of further independent confirmation, are quite encouraging. In addition to demonstrating the utility of the MAS in this clinical setting, the positive findings obtained with this objective instrument suggest that even more fruitful results may be obtained through use of projective techniques designed to uncover indices of emotional factors operating at deeper levels in the personality. It is hoped that the intensive program of investigation we have embarked upon will eventually lead to greater psychological understanding of complex relations between maternal psychodynamics during pregnancy and the process of child development.

SUMMARY

The purpose of this research was to compare measures of manifest anxiety obtained during pregnancy and following childbirth, and to relate these anxiety measures to delivery room experiences. In two independent samples of clinic patients, women who were later to experience complications in the delivery room or were to give birth to children with abnormalities obtained significantly higher manifest scores during pregnancy than did women who later had "normal" delivery room records. The results obtained from retesting one of the samples shortly after childbirth showed decreased levels of manifest anxiety both in patients who had undergone normal childbirth and those who had experienced complications or abnormalities. Manifest anxiety scores were still relatively higher in this latter subgroup, but the difference was no longer significant. It was concluded that these findings demonstrate the clinical utility of the Manifest Anxiety Scale, and also suggest that utilization of projective methods in future research may lead to greater psychological understanding of the role of emotional factors in pregnancy and childbirth.

REFERENCES

- BUSS, A. H., WIENER, M., DURKEE, A., & BAER, M.
The measurement of anxiety in clinical situations.
J. consult. Psychol., 1955, 19, 125-129.

- DAVIDS, A., & ERIKSEN, C. W. The relation of manifest anxiety to association productivity and intellectual attainment. *J. consult. Psychol.*, 1955, 19, 219-222.
- ERIKSEN, C. W., & DAVIDS, A. The meaning and clinical validity of the Taylor anxiety scale and the hysteria-psychasthenia scales from the MMPI. *J. abnorm. soc. Psychol.*, 1955, 50, 135-137.
- FARBER, I. W., & SPENCE, K. W. Complex learning and conditioning as a function of anxiety. *J. exp. Psychol.*, 1953, 45, 120-125.
- GLESER, GOLDINE, & ULETT, G. The Saslow Screening Test as a measure of anxiety-proneness. *J. clin. Psychol.*, 1952, 8, 279-283.
- HILGARD, E. R., JONES, L. V., & KAPLAN, S. J. Conditioned discrimination as related to anxiety. *J. exp. Psychol.*, 1951, 42, 94-99.
- HOYT, D. P., & MAGOON, T. M. A validation study of the Taylor Manifest Anxiety Scale. *J. clin. Psychol.*, 1954, 10, 357-361.
- KENDALL, E. The validity of Taylor's Manifest Anxiety Scale. *J. consult. Psychol.*, 1954, 18, 429-432.
- LUCAS, J. D. The interactive effects of anxiety, failure, and intraserial duplication. *Amer. J. Psychol.*, 1952, 65, 54-66.
- MONTAGUE, E. K. Role of anxiety in serial rote learning. *J. exp. Psychol.*, 1953, 45, 91-96.
- SMITH, W., POWELL, ELIZABETH K., & ROSS, R. Manifest anxiety and food aversions. *J. abnorm. soc. Psychol.*, 1955, 50, 101-104.
- SPENCE, K. W., & FARBER, I. E. Conditioning and extinction as a function of anxiety. *J. exp. Psychol.*, 1953, 45, 116-119.
- SPENCE, K. W., & TAYLOR, JANET A. Anxiety and strength of the UCS as determiners of the amount of eyelid conditioning. *J. exp. Psychol.*, 1951, 42, 183-188.
- SPENCE, K. W., & TAYLOR, JANET A. The relation of the conditioned response strength to anxiety in normal, neurotic, and psychotic subjects. *J. exp. Psychol.*, 1953, 45, 265-272.
- STOTT, D. H. Physical and mental handicaps following a disturbed pregnancy. *Lancet*, 1957, 1, 1006-1012.
- TAYLOR, JANET A. The relationship of anxiety to the conditioned eyelid response. *J. exp. Psychol.*, 1951, 41, 81-92.
- TAYLOR, JANET A. A personality scale of manifest anxiety. *J. abnorm. soc. Psychol.*, 1953, 48, 285-290.
- TAYLOR, JANET A. Drive theory and manifest anxiety. *Psychol. Bull.*, 1956, 53, 303-320.
- TAYLOR, JANET A., & SPENCE, K. W. The relationship of anxiety level to performance in serial learning. *J. exp. Psychol.*, 1952, 44, 61-64.
- THOMPSON, W. R. Influence of prenatal maternal anxiety on emotionality in young rats. *Science*, 1957, 125, 698-699.
- WESLEY, ELIZABETH L. Perseverative behavior in a concept-formation task as a function of manifest anxiety and rigidity. *J. abnorm. soc. Psychol.*, 1953, 48, 129-134.
- WESTROPE, MARTHA. Relations among Rorschach indices, manifest anxiety, and performance under stress. *J. abnorm. soc. Psychol.*, 1953, 48, 515-524.

(Received February 3, 1960)

REPEAT STUDY WITH A PROJECTIVE FILM FOR CHILDREN

MARY R. HAWORTH¹

Michigan State University

Rock-A-Bye, Baby (Haworth, 1960; Haworth & Woltmann, 1959) is a projective puppet film which can be shown to groups of children. The film story focuses on a little boy, Casper, and his jealousy of his baby sister. When left to baby-sit, he begs the witch to help him get rid of the baby. She puts a spell on the milk; mother returns and rushes the baby to the hospital. Casper is filled with remorse, recalls the witch, and finally kills her. Thus the spell is broken, the baby's health is restored, Casper's guilt is resolved, and his parents reassure him of their love by a gift of strawberry ice cream. Woltmann (1951) gives the complete script of the play, as well as the rationale for the use of puppets in projective devices for children.

The film is shown to entire classes, divided into groups of 10 to 15 children per showing. Responses are first secured halfway through the showing when the film is stopped and each child in the group is invited to finish the story. After the rest of the film is shown, each child is asked, individually, a standard set of questions in terms of Casper: what he thought of his parents and of the witch, how he felt when the baby got sick, whether he should be punished for what he did, what he should tell his mother, and how he felt when the baby got well. The child is also asked what part he, himself, liked best and which character he would like to be.

The film was originally administered to 244 children, from nursery school through fifth grade, as reported by Haworth (1957). A scoring scheme (Haworth & Woltmann, 1959)

¹ The author is indebted to the principals and teachers who cooperated in the project, and to Mary Grummon, Ruth Karlake, and James Mathie who assisted in interviewing the children and scoring the protocols.

was devised based on patterns of deviant responses² given to the standard questions and in the group discussion during the showing. The following indices emerged as representing dimensions of personality that appear to be tapped by this particular film: Identification, Jealousy (sibling rivalry), Aggression toward Parents, Guilt (masturbatory), Anxiety (castration), and Obsessive Trends.

The film has subsequently been shown to a new sample of 257 children (kindergarten, first, and second grades) in order to ascertain whether similar proportions of children would score high on the various indices, and whether the developmental progressions which appeared to be demonstrated in the earlier study would be substantiated in the second sample. A cross-validation analysis was planned for the two indices (Guilt and Jealousy) for which criterion groups can be selected from the samples.³ One further aspect of the present study is concerned with scoring reliability.

THE SAMPLES

Table 1 shows the distribution of children, by grades, in the first sample (A, Pennsylvania) and in the second sample (B, Michigan).⁴

Sample A was rather heavily weighted toward the upper professional levels since 95 of the 244 children were from school areas serving predominantly university faculty and professional and managerial groups. The remaining 149 children were drawn from a small

² Deviant responses are those given by less than 10% of a particular age or sex grouping.

³ A subsequent report will be concerned with a validation study in which 15 children who scored high on the Obsessive Index, or on both the Guilt and Anxiety Indices, were matched with 15 low scoring children and given a battery of individual projective tests.

⁴ Discussion of the nursery school group has been omitted as many of the responses were too meager to be scored.

TABLE 1
DISTRIBUTION OF SAMPLES BY GRADES

Sample	Nursery School	Kindergarten	1	2	3	4	5	Total
A	40	—	112	—	45	—	47	244
B	—	86	124	47	—	—	—	257

suburban community representing all occupational levels. In Sample B, half (128) were attending a school which serves the entire range of occupational levels, while 129 children came from a marginal district of predominantly lower class families.

As the original study was particularly concerned with the responses of the large first grade group, approximately the same number of first graders were secured in the second sample for comparative purposes. The groups will hereafter be referred to by number and letter, e.g., 1-A indicates the first grade of Sample A; K-B, kindergarten of Sample B.

RESULTS

Index Scores and Developmental Progressions

A comparison of high scores and deviant identification choices⁵ made by the two first grade samples revealed only one substantial difference: significantly more children of 1-A made deviant identification choices ($\chi^2 = 3.805$; $p = .051$). The specific item that accounted for most of this difference was identification with the opposite-sex parent, with this choice being made more often by 1-A than by 1-B children.

Figure 1 demonstrates the otherwise close correspondence between the two first grades, and includes the fifth grade curve for comparison of the incidence of high scores on each index at different ages.

Developmental progressions for each dimension are shown in Figures 2 and 3. Aggression, Guilt, and Anxiety show congruent curves (Figure 2) with a fairly large incidence of high scores in the early grades and a decided drop occurring between second and third

⁵ Tabulations were made of only the five identification choices which are *always* deviant, as distinguished from another category of choices which are deviant at certain ages or for a specific sex. Each of the other indices requires a specified number of responses for a high score, except Aggression to Parents. For purposes of the present study, the use of even one "aggressive" response is considered a high score.

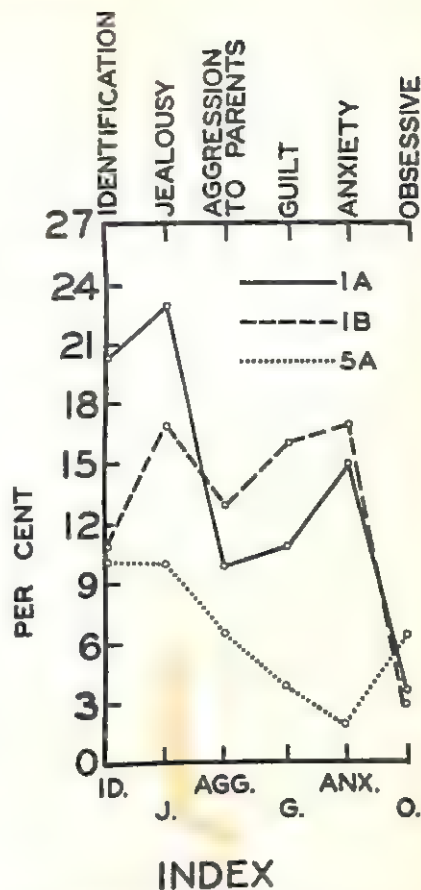


FIG. 1. Percentages of two first grades (1-A, 1-B) and fifth grade (5-A) scoring high on each index.

grades. Figure 3 shows the three indices which maintain fairly constant levels in the later years. Jealousy and Identification start high and remain stabilized at the second grade level, while the Obsessive trends show a constant and much lower level throughout the age range under study.

Cross-Validation of the Guilt Index

The Guilt Index was originally derived (Haworth, 1957) from patterns of deviant responses given by 10 of the 12 children in the 1-A group who had been observed to engage in autoerotic practices (masturbation or thumb sucking) either during the film showing or the inquiry period. Similar response patterns were given by only 2 of the 100 "nonautoerotic" children.⁶ The seven items in

⁶ The Guilt Index did not prove to be applicable to the third and fifth grades. Only 2 (out of the

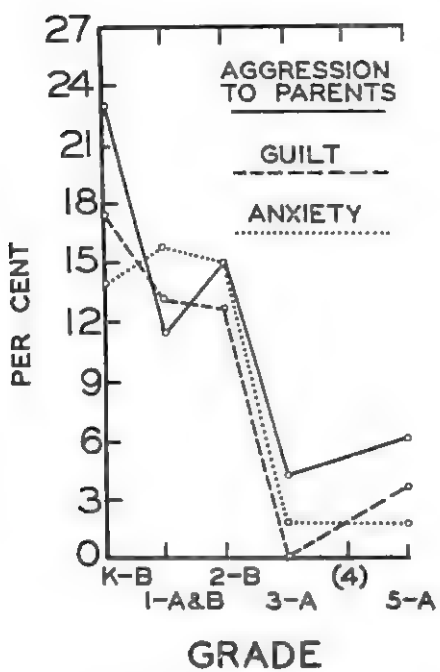


FIG. 2. Percentages of high scores on Aggression to Parents, Guilt, and Anxiety Indices from kindergarten through fifth grade. (Scores are averaged for two first grades.)

the index relate to: (a) mother not knowing what went on, (b) Casper being sent to bed as punishment, (c) Casper feeling rejected by father, (d) Casper being very ashamed or resolving to make amends, and references to (e) the baby stinking, (f) Casper or the baby being in the water, (g) the kissing scenes. Because of the guilt tinged aspect of most of these responses, it would appear that children who respond with high scores (i.e., at least two of the seven items) may be those who not only engage in autoerotic acts but who also feel guilty for so doing. If such responses were also given by the "autoerotic" children in the new sample, considerable validity would be demonstrated for this index.

No statistical test was performed on the 1-A group since this was an ad hoc approach. In the present study, it was predicted that more autoerotic (AE) than nonautoerotic (non-AE) children would score high on the index.

combined total of 92) children received high scores, and neither of these was one of the four observed "autoerotic" cases in the two grades.

Table 2 shows the distribution of high scores (two or more items) as contrasted to low scores (one guilt item or none at all). The predictions were upheld, with significantly more AE than non-AE children receiving high guilt scores; this difference was especially marked in kindergarten and first grade.

The original criterion (Haworth, 1957) for inclusion of a response as an item in the index stated that its incidence in the AE group ($N = 12$) must be at least one-third of its total incidence for all 112 children of the 1-A sample. Actually, for five of the seven items, at least one-half of the responses came from the small AE group. Table 3 shows the distribution of guilt responses in the three grades of Sample B to be quite similar to that of the original 1-A sample from which the index was drawn. It can be seen that, irrespective of high scores, the AE children make more use of the guilt items than do non-AE children, so that the original criterion was still met in all but two instances. (These involved Item No. 5 which was given only once in K-B and in 1-B, and by non-AE children in both

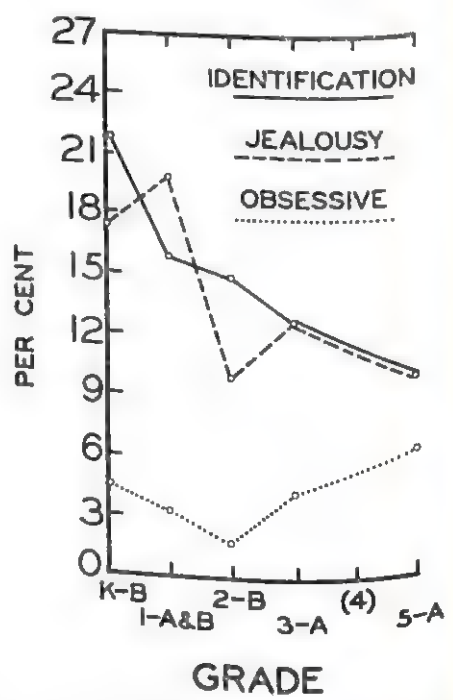


FIG. 3. Percentages of Deviant Identifications and of high scores on Jealousy and Obsessive Indices from kindergarten through fifth grade. (Scores are averaged for two first grades.)

cases.) The criterion was exceeded by five items in K-B, four items in 1-B, and three items in 2-B.

Cross-Validation of the Jealousy Index

In the original study (Haworth, 1957) the items of the Jealousy Index were selected on an a priori basis from responses of the 1-A group which appeared to indicate sibling rivalry. The 11 items of this index include: (a) one response of Casper being jealous of attention given to the baby; (b) a minimum of two uses by the subject of slips of tongue, evasions or personal references; aggression against the baby expressed openly (c) while the film was being shown, (d) in the half-show discussion, or (e-j) in answer to any of six specified inquiry questions. For boys, (k) choosing to be the baby is an additional item. A high score consists of any three of the above responses.

The 1-A sample was divided into two groups on the basis of sibling status, with oldest + middle children in one group, and youngest + "only" children in the other group. Significantly more of the former group scored high on the Jealousy Index, and the difference was largely due to the boys' responses in each grouping. Within the oldest + middle grouping, significantly more oldest than middle children received high scores.

A similar analysis of the 1-B sample (as well as K-B and 2-B) revealed no differences between the various groupings: oldest + middle vs. youngest + only, oldest + middle boys vs. youngest + only boys, or oldest vs. middle children. The total incidence of high jealousy scores was also less for 1-B (16.9%) than for 1-A (23.2%), but this difference was

TABLE 2

INCIDENCE OF HIGH AND LOW GUILT SCORES IN AUTOEROTIC AND NONAUTOEROTIC GROUPS (SAMPLE B)

Grades	N	Guilt Scores			
		High	Low	χ^2	p
K-B					
Autoerotic	28	11	17	11.61 ^a	<.001
Nonautoerotic	58	4	45		
1-B					
Autoerotic	32	13	19	16.76 ^a	<.001
Nonautoerotic	92	7	85		
2-B					
Autoerotic	16	5	11		.013 ^b
Nonautoerotic	31	1	30		

^a Corrected for continuity.

^b Fisher exact probability test.

not significant. There were some indications that differences in family size, ordinal position, or socioeconomic status might be responsible for the lack of replication in Sample B. Much larger samples would be required to secure enough high scoring cases for an analysis of these multiple variables.

Reliability

Three judges scored a group of 24 protocols pulled at random from the B sample. Inter-scoring reliability was computed for the four main indices: Jealousy, Guilt, Anxiety, and Obsessive Trends. (No reliability study seems necessary for Identification choice since quite objective criteria can be applied; Aggression

TABLE 3
INCIDENCE OF GUILT INDEX ITEMS IN AUTOEROTIC (AE) AND
NONAUTOEROTIC (NON-AE) GROUPS

Guilt Items	1-A		K-B		1-B		2-B	
	AE (N=12)	non-AE (N=100)	AE (N=28)	non-AE (N=58)	AE (N=32)	non-AE (N=92)	AE (N=16)	non-AE (N=31)
1. Mother didn't know	4	0	2	1	4	1	1	0
2. Bed or sleep	9	10	11	10	14	23	3	7
3. Father rejects Casper	4	2	6	4	3	2	4	0
4. Guilt/restitution	4	6	7	5	10	10	6	2
5. Baby stinks	2	2	0	1	0	1	0	0
6. Casper or baby in water	2	5	3	0	0	0	0	0
7. Kissing	2	2	1	0	3	0	0	0

to Parents is used qualitatively and has not been set up in terms of number of responses necessary for a "high" score.) The number of checks given to each item of each index was compared for each of the three pairs of judges. The Rulon formula, with the Spearman-Brown correction, yielded the following reliability coefficients for each index:

Jealousy: .95, .93, .95; average = .94
 Guilt: .78, .88, .82; average = .83
 Anxiety: .92, .91, .92; average = .92
 Obsessive: .94, .87, .92; average = .91

While the overall reliability of scorers is quite satisfactory, the lower agreements on the Guilt Index were examined for possible causes. It was found that most of the discrepancies between judges occurred as the result of neglecting to check, under the item "Father rejects Casper," those responses in which the father is mentioned specifically as the punisher. The directions have subsequently been clarified to call attention to this objective point.

DISCUSSION

A replication of the film test has revealed no appreciable differences between the two large first grade samples, except in the area of deviant identification choices, and in the sibling status (but not the incidence) of children responding on the Jealousy Index. The repeat study has also confirmed the earlier impression that developmental progressions occur in some areas while plateaus are maintained in other dimensions. The fact that similar and congruent results were obtained between two samples differing in location and socioeconomic composition demonstrates a certain amount of construct validity for the test. To put it differently, if marked differences and discrepancies had been found, then very little confidence could be put in this instrument as a method of personality assessment.

As was expected on the basis of original findings, the younger children express more outspoken aggression toward parents than do older children, and they also score higher on measures of guilt and anxiety. There appears to be no reason to abandon the earlier hypothesis (Haworth, 1957) that this film does

pinpoint certain problem areas related to the oedipal period. In view of the decided drop in incidence of guilt and anxiety between the ages of 7 and 8 (by which time the latency period is presumed to be well underway), it still seems, as originally suggested, that the guilt measured by the film test is associated with masturbation and other autoerotic acts. (The postulated relationship between anxiety and castration fears is currently being studied via other projective techniques.) It can only be speculated whether the slight trend upward of the obsessive scores between the second and fifth grades may indicate an increasing incidence at still later ages. Fenichel (1945) sees an increase in obsessive reactions and compulsive rituals during the latency period as defenses become strengthened against the instinctual impulses. The curves in Figures 2 and 3 may possibly be a graphic representation of the repression of erotic drives and the development of defense mechanisms.

If identification patterns are laid down during the oedipal period, the incidence of deviant identifications should remain at fairly stable levels throughout the age range studied. This was found to be the case. On the basis of the film responses it would appear that jealous reactions, once established, also do not decline in the early latency period. The threat to the ego is undoubtedly not as great in this area as in those more closely linked to the oedipal situation. Consequently there would be less need to repress or defend. In some instances jealousy toward siblings may even be serving as a substitute outlet for unacceptable feelings originally directed toward the parents.

The one significant difference between the two first grade samples—namely, deviant identification—may possibly be attributable to differences in the socioeconomic status of the two groups. The item responsible for most of this difference was the choice of the opposite-sex parent by more children from the higher status group. This finding is consistent with that of Rabban (1950) who showed sex-role identification to be more clearly defined, and at an earlier age, for lower class children.

With respect to the Guilt Index, as has been previously pointed out (Haworth, 1957),

it is not to be expected that all autoerotic children would feel guilty. Nevertheless, the fact that a repeat study still shows large proportions of them giving a specific cluster of responses suggests that dynamic factors are being tapped by the index, namely, conflicts between instinctual drives and conformity to parental standards. The validity of the Guilt Index for children from kindergarten through second grade has also been demonstrated by the consistently significant differences between the number of high scores in the autoerotic, as contrasted to the nonautoerotic, groups.

In spite of consistent findings on the Jealousy Index with respect to the frequency of children receiving high scores, the sex and sib-status distribution of the scores was not upheld in the second sample. It appears that high scores may be measuring attitudes to either older or younger siblings. In view of the equivocal findings, caution should be exercised in the interpretation of this index, especially if it is the only high score in a protocol. In combination with high scores on other indices, it may provide useful supplementary data for diagnostic purposes.

SUMMARY AND CONCLUSIONS

The projective puppet film, *Rock-A-Bye, Baby*, was originally shown to 244 children from nursery school through fifth grade. The film has subsequently been shown to 257 children from kindergarten through second grade. The two large first grade samples showed close correspondence with respect to incidence of deviant scores on all measures except Identification. The consistent developmental progressions from grade to grade, within and

between samples, demonstrate construct validity for the instrument.

Two indices could be cross-validated by means of criterion groups within each sample. The Guilt Index showed the predicted significant differences between autoerotic and nonautoerotic groups in all three grades of the new sample. Differences between sibling groupings were not upheld on the Jealousy Index.

Since adequate interscorer reliability has been demonstrated for the instrument, and generally consistent kinds of data have been secured in a replicated study, confidence can be placed in this technique as a group screening device in the personality assessment of early latency children.

REFERENCES

- FENICHEL, O. *The psychoanalytic theory of neurosis*. New York: Norton, 1945.
- HAWORTH, MARY R. The use of a filmed puppet show as a group projective technique for children. *Genet. psychol. Monogr.*, 1957, 56, 257-296.
- HAWORTH, MARY R. Films as a group technique. In A. I. Rabin & Mary R. Haworth (Eds.), *Projective techniques with children*. New York: Grune & Stratton, 1960. Pp. 177-190.
- HAWORTH, MARY R., & WOLTMANN, A. G. *Rock-A-Bye, Baby: A group projective test for children*. (Manual and Film) University Park, Pa.: Psychological Cinema Register, 1959.
- RABAN, M. Sex-role identification in young children in two diverse social groups. *Genet. psychol. Monogr.*, 1950, 42, 81-158.
- WOLTMANN, A. G. The use of puppetry as a projective method in therapy. In H. H. Anderson & Gladys L. Anderson (Eds.), *An introduction to projective techniques*. New York: Prentice-Hall, 1951. Pp. 606-638.

(Received February 4, 1960)

LENGTH OF THERAPY IN RELATION TO OUTCOME AND CHANGE IN PERSONAL INTEGRATION¹

DESMOND S. CARTWRIGHT, RICHARD J. ROBERTSON,
DONALD W. FISKE, AND WILLIAM L. KIRTNER²

University of Chicago

Standal and van der Veen (1957) have recently suggested that number of interviews constitutes an important variable for study in research on psychotherapy. Their argument lay especially in the demonstration that of several measures of progress in therapy, based upon counselor judgments, a measure of change in personal integration of the client was not only the most important clinically and theoretically, but also showed the highest linear correlation with log number of interviews.

If it were substantiated that a very high correlation between length of therapy and change in personal integration exists, it would be an important finding, indeed; for it might be possible to employ number of interviews as a dependent variable of exceptional reliability which nevertheless has critical implications for personality change. There can be no doubt that the finding of dependent variables which have both high reliability and high validity and also relevant clinical implications is among the most important tasks of psychotherapy research today.

Therefore, it seems important to confirm this earlier finding and to attempt to determine the most valid procedure, among possible alternatives, for obtaining a measure of personal integration derived from counselor judgments. To illustrate the alternatives: we note that the counselors who made judgments of change in personal integration for the Standal and van der Veen (1957) study did

so at the end of their series of contacts with clients. In this procedure the counselors were asked to rate the integration of each client at termination and at the same time to scan their long-term memories for a rating of the initial level of integration. A nine-point scale ranging from "highly disorganized or defensively organized" (1) to "optimally integrated" (9) was used and *change in integration* was defined as the arithmetic difference between initial and terminal scores.

If we consider the counselor's thoughts in making such ratings it seems reasonable to suspect a certain bias resulting from sheer lapse of time involved. The longer the time of acquaintance experienced by the counselor, the greater his tendency to underestimate the initial level of integration. For suppose a counselor were really just guessing about the level of integration of his client a long time ago, his thought might well go something like this: "The client has been with me a very long time so he must have been in rather poor shape to begin with."

An alternative procedure would be to eliminate this sort of possible bias by obtaining judgments of the level of integration actually perceived by the counselor at both beginning and end points. Since counselor judgments are not only the most frequently employed criterion measures of progress, but also the most available, the present study was undertaken to compare the two rating procedures when applied to data comparable to that of Standal and van der Veen (1957). In addition it was hoped to tease out some of the differences, if any, between a measure of therapy length based upon number of interviews as against one based solely on number of weeks.

¹ The study was supported by funds from the Ford Foundation Psychotherapy Research Fund, granted to the Counseling Center, University of Chicago.

² D. S. Cartwright now at the University of Colorado; W. L. Kirtner now at the California Institute of Technology.

SUBJECTS AND PROCEDURE

From a single large research block of cases at the Counseling Center, University of Chicago, 87 clients had terminated therapy at the time the present study was undertaken. (Omitted were 6 cases which had started in the block but were still in therapy.) All clients had been seen during the period 1956-59 by client centered therapists. These therapists included both males and females, and their experience levels ranged from 1 to 12 years. There were 52 male clients and 35 females. Also, 52 of the clients were students, 35 were not. The mean age of the clients was 28.5 years ($SD = 7.6$), and they had been in therapy for a mean number of 29.5 interviews ($SD = 28.1$), and for a mean number of 31.9 weeks ($SD = 22.5$).

Two measures of length of therapy were taken: the exact number of interviews and the rounded number of weeks. Counselors were asked to make ratings on their clients immediately after the last interview and also immediately after the last interview. In the large majority of cases, ratings were dated between 1 and 3 days after the relevant interview. The number of weeks was computed from the number of days lapsing between initial and final dates of ratings.

Two measures of integration movement were used. The first measure was taken only at the end of therapy, thus involving the counselor's long-term memory. The counselor was asked to answer two questions. The first was: "What change has there been in the client's feelings toward himself?" Four response alternatives were provided, ranging from "more discontented" through "much more contented." Scores of 1 through 4, respectively, were assigned. The second question was: "How much change in the client as a person has occurred since he started counseling?" Four response alternatives were provided, ranging from "not changed" through "changed a good deal." Scores of 1 through 4, respectively, were assigned. The sum of the scores on these two questions constituted the first measure of integration change. It will be called the posttherapy estimate of change in integration (PECI).

The second measure of integration change was a difference score between two ratings, one made after the first interview, one made after the final interview. Thus, on each rating occasion, only the counselor's short-term memory was involved. Ratings were made on a 10-point scale, ranging from "most extreme maladjustment" through "optimal adjustment (fully functioning, optimal maturity)." A score of 1 was assigned to the most maladjusted end, a score of 10 to the optimal adjustment end. Using this scale, the counselor was asked to indicate his estimate of the client's present psychological adjustment. The score for his estimate after the initial interview was subtracted from the score for his estimate after the final interview to yield the second measure of change. This second measure will be called the difference measure of change in integration (DMCI).

In addition to the above measures, the counselor's nine-point rating of success of the therapy was taken for this research. The scale has been used for many years at the Counseling Center, and was employed also by Standal and van der Veen (1957). The score of 9 means marked success.

The reliability and validity of the measures PECI and DMCI are not known independently of the present study. However, it will be shown in the results below that both have strong correlations with the success rating and with each other. The success rating scale has previously been shown to have substantial reliability and validity (Cartwright, 1955).

RESULTS

The comparability between the samples studied by Standal and van der Veen (1957) and the present writers is very good. Notably, the mean length of therapy was 30.7 interviews ($SD = 32.5$) in the former study, 29.5 ($SD = 28.1$) in the present study. Both samples have a slightly greater proportion of male than female clients, and of student than community clients. For both studies male and female therapists were employed.

The data basic to replicating the major results of Standal and van der Veen (1957, p. 9) are included in Table 1, which shows intercorrelations of all the measures taken in the present study.

First, both PECI and DMCI correlate positively and significantly ($p < .001$ and $p < .01$, respectively) with log number of interviews. Thus, the first major conclusion of Standal and van der Veen (1957), that "Change in level of personal integration . . . has a moderate linear relationship with log case length" (p. 9), is supported.

TABLE 1

INTERCORRELATIONS OF TWO MEASURES OF LENGTH OF THERAPY, TWO MEASURES OF CHANGE IN PERSONAL INTEGRATION, AND A RATING OF SUCCESS
($N = 87$)

	Log Number of Weeks	Log Number of Inter- views	PECI	DMCI
Log Number Interviews	.85			
PECI	.22	.36		
DMCI	.10	.29	.64	
Success	.29	.49	.72	.68

Note.—For $r = .35$, $p < .001$; $r = .28$, $p < .01$; $r = .21$, $p < .05$.

Second, the success rating correlates positively and significantly ($p < .001$) with log number of interviews. This finding accords with that of Standal and van der Veen (1957, p. 6), but the relative sizes of the correlations for success rating and for change in personal integration with log number of interviews differ in the two studies. Whereas Standal and van der Veen found the Pearson correlation for change in personal integration to be .58, and that for success rating to be .37, in the present study the order is reversed for both measures of change in personal integration as compared with success rating. Thus, the second major conclusion of Standal and van der Veen (1957), that "Change in level of personal integration is more highly related to case length than change or outcome on other important case variables" (p. 9), is not supported. This finding also lends no support to their fourth major conclusion, that "With respect to actual amount of therapy, change in personal integration may be more important than rated success or other case variables" (p. 9). At this time, one can say only that length of therapy is positively related to several measures of outcome or change.

It should be noted that the above results hold for both a measure of change in personal integration which relies to some extent on the counselor's long-term memory (PECI) and a measure of change which does not rely on long-term memory (DMCI). Examination of the correlations between log number of weeks and the three case variables in Table 1 shows that the two measures taken only at posttherapy (PECI and the success rating) have significant positive correlations, while the difference measure which does not rely on long-term memory has a nonsignificant correlation. Since it makes little sense to partial out number of weeks from number of interviews the evidence in Table 1 must be taken as it stands to suggest that sheer length of acquaintance does have some influence on the counselor's ratings when these ratings involve his use of long-term memory.

The question arises whether it is possible to show somewhat more conclusively the postulated effects of long-term memory on the counselor ratings of change made at the end of therapy. The first thing that may be noted

from the reliability data presented by Standal and van der Veen (1957, p. 5) is that the rate-rerate reliability coefficient for personal integration at the beginning of therapy, as rated at the end of therapy, was .50 (not significant); while the comparable coefficient for the termination of therapy was .68 (significant at the .05 level). It is also noteworthy that in their discussion of reliability they reported that 34 months later, certain counselors could not remember well enough to make reratings on certain items. The present concern is whether the counselors at the time of their first rating could remember enough about the beginning of therapy to make valid ratings. It was suggested above that with long cases, the counselors might have been sufficiently hazy in their long-term memory to be rating essentially on a guessing basis with a bias toward underestimating the level of integration shown by clients at the beginning of therapy. To examine this issue, the original data for the 72 clients reported on by Standal and van der Veen (1957) in regard to ratings of personal integration were re-examined along with the ratings on DMCI for the present sample. These authors report a Pearson correlation of .67 between the success rating and change on personal integration. Table 1 shows the Pearson correlation of success rating and DMCI to be .68.

The two scales are highly comparable. They have closely similar wording. The first has 9 steps, the second has 10. Further, it was found that the variances were not significantly different. Inspection of the distributions and of the wording for the bottom point suggested that the scales could be considered essentially equivalent if the unused bottom step of the 10-point scale was dropped and the other steps renumbered accordingly.

Table 2 summarizes the comparisons between the ratings for the two studies when the scale used in the present study is treated as a nine-point scale.

Table 2 indicates that for the two samples, the difference between the posttherapy ratings is not significant while the difference between the pretherapy ratings is highly significant. (Even if the latter difference is reduced by .31, the amount of the posttherapy difference, the t -value is still very high—3.85.) Thus, for

TABLE 2

COMPARISON OF MEAN RATINGS OF PERSONAL INTEGRATION FOR TWO STUDIES

Study	Period Rated	N	M	SD	t
Standal & van der Veen	Beginning	72	3.1	1.4	5.22**
Present Study	Beginning	87	4.3	1.4	
Standal & van der Veen	End	72	5.4	1.7	1.22*
Present Study	End	87	5.7	1.4	

* $p < .20$.
 ** $p < .001$.

comparable samples, there is no difference for the posttherapy ratings which were based on the short-term memory. For the ratings of integration at the beginning of therapy however, the mean rating of integration is significantly lower for the sample studied by Standal and van der Veen (1957). This result is in accord with the expectation that counselors relying on long-term memory would tend to underestimate the degree of personal integration shown by their clients at the beginning of therapy.

DISCUSSION

The absolute size of the correlations obtained in this study between log number of interviews and measures of change in personal integration was not very great, even though the latter judgments may be influenced by knowledge of the former. While it does not seem likely that number of interviews can be used as a clinically meaningful dependent variable on its own, it does bear useful relations to a number of important measures of change taken from counselor judgments, and these relations appear to be quite stable over the two samples studied. It is clear from the present findings, however, that considerable caution must be exercised when employing counselor judgments to obtain such estimates or estimates of any variable. In particular it appears important to pay careful attention to the conditions under which counselor judgments are obtained, especially in regard to the time span over which

they are called upon to exercise their memories.

The question of whether log number of weeks or log number of interviews is the better measure of length of therapy cannot be given a general answer from the present study. So far as the evidence does go, it appears that log number of interviews shows the higher correlations with measures of change in personal integration and success of therapy when these are taken from counselor judgments. However, it also seems that a spurious length-of-acquaintance factor may be contributing to those higher correlations when the measures are taken from counselor judgments made only at the termination of therapy. All in all, the best procedure at the present time would seem to be offered by the use of log number of interviews in conjunction with judgments made both at the beginning and at the termination of therapy.

SUMMARY

Data for 87 clients seen by client centered counselors were examined in order to replicate certain analyses made by Standal and van der Veen (1957) on a similar sample. It was confirmed that counselor rating of movement on personal integration bears a linear relationship to log number of interviews. In contrast to the earlier results, the present study found that the counselor success rating had a higher correlation with length of therapy than did rated movement on personal integration. An alternative measure of length of therapy was also employed in the present study, namely log number of weeks. Correlations with movement and success were uniformly smaller for this alternative measure of length.

Memory factors influencing the counselors' judgments were examined by use of two measures of change in personal integration, one calling upon the counselor to rate change at the end of therapy only, the other calling upon him to rate the level of integration he sees in the client after the first interview and after the final interview, change being calculated from the difference between the initial and final ratings. The results showed that, when counselors' ratings of change are made

only after termination of therapy, they are influenced by the sheer length of acquaintance with the client. It was also hypothesized that in the Standal and van der Veen (1957) study, counselors who, after the termination of therapy, rated the initial level of personal integration of the client would have been operating on such long-term memory as to involve considerable guesswork coupled with a bias to underestimate the client's initial level of integration. This hypothesis was tested by comparing the mean rating of initial integration in the earlier study with the mean rating of initial integration in the present study when counselors were rating from short-term memory. The result supported the hypothesis.

It was concluded that the use of log number of interviews together with judgments made both at the beginning of therapy and at the termination appears to be the best present procedure for examining the relations between length of therapy and case variables obtained from counselor judgments.

REFERENCES

- CARTWRIGHT, D. S. Success in psychotherapy as a function of certain actuarial variables. *J. consult. Psychol.*, 1955, 19, 357-363.
- STANDAL, S. W., & VAN DER VEEN, F. Length of therapy in relation to counselor estimates of personal integration and other case variables. *J. consult. Psychol.*, 1957, 21, 1-9.

(Received February 12, 1960)

BENDER-GESTALT FIGURE ROTATIONS: A STIMULUS FACTOR

RICHARD M. GRIFFITH AND VIVIAN H. TAYLOR

Veterans Administration Hospital, Lexington, Kentucky

Hannah (1958) modified the Bender Visual Motor Gestalt Test (BG) by rotating the designs through 90 degrees on their rectangular cards, presenting the design to the subject as before. With this new set of cards patients produced fewer rotations of figures, presumably because the longer axis of the card now corresponded to the longer axis of the paper. However, his statistically significant difference was due to a few of his controls producing multiple rotations; examining his statistics it becomes evident that just as many patients in one group as in the other rotated at least one figure (8 out of 36 in each case). The present study is essentially a replication of his; however, instead of redesigning the cards, a comparable effect was attained through the expedient of rotating the paper, card and paper being thus oriented lengthwise left-to-right instead of up-and-down as was his.

Examiners within a large neuropsychiatric hospital were asked to rotate the tablet when administering the BG. Habits being hard to break, not all did so. Those who did not comply unwittingly collected a "control" group of records, which, as it turned out, matched the experimental as to diagnosis. As the psychological reports crossed the secretary's desk rotations were noted, the study continuing over a 10-month period. An angular displacement of at least 45 degrees in a recognizable figure was the criterion for rotation.

Fifty-six "tablet-turned" records were obtained, 157 conventional ones. Under the modified conditions 12.5% of the records had one or more figure rotations vs. 29.3% under the conventional, the two proportions differing significantly at the .02 level (one-tailed

test). A chi square between the distribution of diagnoses in the two groups (five major diagnostic categories being considered) was small—0.650 for 4 degrees of freedom—permitting the conclusion that the groups were well matched according to diagnosis.

The 29.3% rotations in the standard records were unaccountably higher than the 22.8 previously determined from approximately 1,000 records in the files of the same hospital (Griffith & Taylor, 1960). After a chi square test had shown that there had been no statistically significant shift in diagnoses, all the data collected under standard conditions were combined for a total number of 1,152 tests—23.5% with one or more figure rotations. This 23.5% differed just at the .05 level of statistical significance from the 12.5% of records with rotations in the unconventional, tablet-turned group (one-tailed test).

Hannah's results would seem to be confirmed. It may be concluded that many rotations are caused by the patient orienting the design to the major axis of the paper in the same relation it bears to the major axis of the card, even though to do so involves actually turning the design in relation to himself. The results fit into the pattern of investigations begun by Shapiro (see Williams, Lubin, Giesekeing, & Rubinstein, 1956) which relate the phenomena of rotations of both block designs and BG figures to stimulus properties of figure and ground. However, it should be pointed out that however successful we may be in pinpointing the stimulus variables which influence rotations, rotations do not thereby lose their diagnostic significance; as long as different diagnostic groups are influenced differently by the stimulus con-

ditions as they seem to be (Griffith & Taylor, 1960) the rotation will still have diagnostic significance.

To sum up, it was confirmed, through a replication of a previous study, that many of the rotations of the Bender-Gestalt figures may be attributed to the accidental circumstance that the long axis of the test card is oriented at 90 degrees to the long axis of the paper upon which the figure is usually drawn.

REFERENCES

- GRIFFITH, R. M., & TAYLOR, V. H. Incidence of Bender-Gestalt figure rotations. *J. consult. Psychol.*, 1960, 24, 189-190.
- HANNAH, L. D. Causative factors in the production of rotations on the Bender-Gestalt designs. *J. consult. Psychol.*, 1958, 22, 398-399.
- WILLIAMS, H. L., LUBIN, A., GIESEKING, C., & RUBINSTEIN, I. The relation of brain injury and visual perception to block design rotation. *J. consult. Psychol.*, 1956, 20, 275-280.

(Received January 28, 1960)

BRIEF REPORTS

THE SOCIAL DESIRABILITY SET IN INDIVIDUAL AND GROUPED SELF-RATINGS¹

NORMAN A. MILGRAM AND MALCOLM M. HELPER

Nebraska Psychiatric Institute

High correlations have typically been found between a group's self-ratings on an array of items and the social desirability of those items. Taylor (1959) has challenged the conclusion that each S's self-ratings reflect only his desire to produce a favorable self-picture.

The present study replicates Taylor's design in comparing individual and grouped data, but uses normal Ss, a shorter time interval between ratings, different instructions, and a different rating instrument—one whose test-retest reliability for self-ratings had been studied.² In addition, it attempts to manipulate the desirability set in individuals by exposing them to a personally relevant ideal prior to obtaining their self-ratings; presumably this exposure would enhance the desirability set.

Eighty incoming freshman male medical students ranked the definitions of the 15 Murray needs given by Edwards (1957) from most to least characteristic of themselves, and in a separate ranking, from most to least characteristic of successful physicians. Forty Ss ranked the items first for themselves and immediately thereafter for physician (Group S-P), while the other 40 followed the reverse sequence (Group P-S).

In Group S-P, where the rating sequence was comparable to Taylor's, the pattern of results was similar to his. Rank-order correlation between average ranks assigned to the items for self and for physician was .89, while the median of the individual correlations was only .63, with

11 of 40 Ss having correlations below .44, the .05 significance level. These results support Taylor's contention that for the self-ratings of a substantial portion of Ss, factors other than social desirability set are operative.

In the group receiving reversed-order instructions (Group P-S), however, the median individual correlation (.85) was significantly higher than that in Group S-P ($p < .01$, median test) and close enough to the correlation based on item means (.95) to suggest that little but the desirability set was operating in these Ss; only two individual correlations in this group fell below .44. Apparently making physician ratings first enhanced the desirability set in the subsequent ratings of self.

That occupying the second position in the instruction sequence modified the *self-ratings* in Group P-S, and not the *physician ratings* in Group S-P, is indicated by an additional finding: self-ratings in Group P-S were more uniform than in Group S-P, while there was no difference in physician ratings. When each S's self-rating was correlated with the mean self-rating for his group, the median rho for Group P-S was .80 and for Group S-P .62, the median test being significant at the .01 level. Physician ratings were higher and equally uniform for P-S and S-P groups, the median rho's being .87 and .85, respectively.

In addition to corroborating Taylor's findings, the present study provides evidence that the desirability set in self-ratings can be enhanced by simply having Ss make desirability ratings first.

REFERENCES

- EDWARDS, A. L. *Manual for the Edwards Personal Preference Schedule*. New York: Psychological Corp., 1957.
TAYLOR, J. B. Social desirability and MMPI performance: The individual case. *J. consult. Psychol.*, 1959, 23, 514-517.

(Received February 11, 1960)

CHOICE DISCRIMINATION IN SCHIZOPHRENIC AND NORMAL SUBJECTS FOR POSITIVE, NEGATIVE, AND NEUTRAL AFFECTIVE STIMULI¹

MILTON TURBINER

Veterans Administration Hospital, Northport, New York

This study was concerned with the effects of three varying affective stimuli on discriminative performance of schizophrenic and normal subjects (Ss). It was hypothesized that the apparent ineffectiveness of schizophrenic Ss in discrimination tasks is related to certain motivational factors within the stimulus situation.

Twenty male schizophrenic Ss and an equal number of normal Ss were used in this study. Three series of pictorial stimuli were selected corresponding to the dimensions of positive, negative, and neutral affective states. Each scene was represented by five pictures. Two of the series consisted of a social situation involving a female figure whose face and hands were clearly in evidence and a three-fourths rear profile view of a young child in the foreground. The third series consisted of a geometric design with the intent that these pictures were to represent a minimal amount of any given affective quality. The series categorized as negative affective contains the theme of reprimand by the central female figure with respect to the child; the positive affective series contains the theme of acceptance and desire for closeness on the part of the woman with

respect to the child. Size, position, and general physical characteristics of the characters were held constant in both affective series, except for a progressive alteration of the facial expression of the central figure and the change in the position of her hands—from those representing a closeness to those representing rebuff.

The instructions required that upon simultaneous presentation of the pairs of stimuli of each series the Ss were to indicate whether the moods expressed in the central figure in both pictures were the same or different. Similar instructions were given for the discrimination of the geometric series. The scores obtained for each S consisted of the frequency with which the S responded "same" to a pair of different pictures and "different" to a pair of identical scenes.

Coincidental with the writing of the dissertation from which this brief report is derived Dunn (1954) published a study based upon a similar hypothesis and research design. The findings of this study are generally consistent with those reported by Dunn (1954). It was found that the performance of the schizophrenic group was less effective in contrast to that of the normals with respect to negative as well as positive affective stimuli discrimination. However, their performance was indistinguishable from that manifested by the normal group with respect to the neutral stimuli. This is a clear indication of a capacity common to both normal and schizophrenic groups, which, under predicted conditions, was not utilized effectively by the schizophrenic group as stimuli conditions changed.

REFERENCE

- DUNN, W. L. Visual discrimination of schizophrenic subjects as a function of stimulus meaning. *J. Pers.*, 1954, 23, 48-64.

(Received April 11, 1960)

¹ This paper is derived from a doctoral dissertation submitted in partial fulfillment of the requirements for the degree of PhD, Boston University Graduate School, 1955. Grateful acknowledgment is due L. J. Reyna and J. V. Gilmore for their help and guidance.

An extended report of this study may be obtained without charge from Milton Turbiner (Box 326, Veterans Administration Hospital; Northport, New York) or for a fee from the American Documentation Institute. Order Document No. 6411 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance \$1.75 for microfilm or \$2.50 for photocopies. Make checks payable to: Chief, Photoduplication Service, Library of Congress.

CLINICAL PERCEPTION OF THE THERAPEUTIC TRANSACTION

BERTRAM R. FORER, NORMAN L. FARBEROW, HERMAN FEIFEL,
MORTIMER M. MEYER, VITA S. SOMMERS, AND RUTH S. TOLMAN

*Veterans Administration Outpatient Clinic, Los Angeles*¹

METHOD

An important facet of the diagnostic and therapeutic work of the clinician is that the data upon which he makes his decisions be both relevant and determinate. In all clinical fields there are data of such universality as to be of essentially no differential significance (Forer, 1959). Diagnosis in internal medicine is not furthered by such facts as the existence of a given organ. Similarly clinical psychologists are minimally assisted in their decisions by knowledge that their patients are human beings, have problems, possess unfulfilled needs, and the like. Clinical practice presupposes recognition and evaluation of characteristics that vary sufficiently among clients to permit the description of uniqueness.

But this is not enough. It is of additional importance that characteristics in which people do vary be relatively stable for a given individual. If, for one client, a particular psychological characteristic were to manifest a temporal variation approaching that of a large sample of individuals at any one time, clinical description would be largely a matter of chance. The same would be true of the individual described by a group of observers whose ratings covered the total population range on a given trait.

In an attempt to clarify some of the language of clinical observation and description and to experiment with suitable research methods for further work, the writers sought to determine in concrete clinical situations: how well a group of trained observers would agree on what they perceived, what kinds of clinical material could be perceived most reliably, and whether clarification of descriptive language and concepts would enhance agreement among judges.

To this end a check-list was developed to include a sample of items which might be manifested in a therapeutic hour. Items were selected which would likely characterize some, but not all, clinical interactions. All items were cast in a form that permitted a rating of "present" or "not present." The number of items changed as the experiment progressed as indicated in Table 1. Items were classified a priori by the judges as either "observational" or "inferential" according to the degree of extrapolation beyond immediate data believed necessary to make the clinical judgment.

Judges were six diplomates in clinical psychology, five of whom had worked together in the same clinic for 8 to 11 years, and one who had been with the group for about 3 years and whose clinical training and background were similar in content and duration. All were joint participants in training, supervision, diagnosis, therapy, seminars, and research in a psychoanalytically oriented Veterans Administration Mental Hygiene Service.

The judges observed 50-minute psychotherapeutic sessions between a patient and a therapist through a one-way screen with a microphone-amplifier hookup. Judges made independent ratings without discussion. Observations consisted of four phases:

A1-3. A male psychology trainee with a neurotic woman patient. They had been working in the observation room for some time, but did not know of the group's observation. There were three weekly sessions followed by a discussion and revision of some items.

B1-3. A female psychiatrist with a schizophrenic male patient. They had been working together for some time in another room and the therapist moved into the observation room at the experimenter's request. There were three sessions followed by discussion and revision of items.

B4-6. The same patient-therapist team. There were three more weekly sessions followed by discussion and more extensive revision and redefinition of items (Table 1).

C1-3. A male psychology trainee with a neurotic male patient whom he had been seeing in the room for some time for supervisory purposes. They knew nothing of the experiment.

¹ Formerly part of Veterans Administration Regional Office, Los Angeles.

TABLE 1
CHECK LIST FOR CLINICAL OBSERVATIONS

1. Therapist was primarily active. (Active: verbal activity)^a
2. Therapist was primarily comfortable. (Comfortable: has sense of ease with the patient. Frame of reference should be our concept of comfort for all therapists.)

Therapist's major method was:

3. Reflective
4. Interpretive
5. Supportive
6. The session seemed to be focused on a problem.
 - (6a. Therapist attempted to focus on a particular problem, either content or dynamic.)
 - (6b. Patient tended to ramble from topic to topic.)
7. There were silences. (Silence: period when no one is talking, 30 seconds or more)
8. If there were silences, they were usually broken by:
 - a. the patient
 - b. the therapist
9. The hour was characterized by resistance.
 - (9a. Patient expressed verbal disagreement with therapist's interpretations, regardless of unconscious factors.)
 - (9b. Patient interrupted therapist more than once.)
 - (9c. Patient was halting in his speech: made pauses before words, incomplete sentences, more than once.)
 - (9d. Patient spoke mostly in generalities.)
 - (9e. Patient used technical psychological terms, more than once.)
10. Patient spoke in a monotone.
11. Patient was fidgety or restless.

Problem areas worked on were:

12. Authority
13. Sex
14. Dependency
15. Work
16. Hostility
17. Emotional control
18. Symptoms
19. Relationships with people^b
20. Patient used gestures. (More than once)
21. Patient brought up a lot of material. (Material: variety or elaboration of content)
22. The material brought up was deep. (Deep: (a) patient meaningfully relates something that is happening present to what has happened in past [content], or (b) produces something with great affect)
23. Patient was experiencing affect.

If experiencing affect, it was:

24. Anger
25. Fear
26. Sadness
27. Anxiety
28. Warmth
29. Patient seems to be rigid. (This applies to fluidity and spontaneity during the hour, whether material is censored or uncensored, etc. Not a judgment of character structure.)
30. Patient seems to show ability to form close relationships.
31. Patient seems capable of insight.
32. Patient seems self-critical.^b
33. (Patient's relationship to the therapist during the hour seemed to be one of positive feelings or rapport. This refers to conscious feelings.)
- 34a. The nature of the transference is predominantly positive.
- 34b. The nature of the transference is predominantly negative.

^a Items in parentheses are clarifications or substitute items created during the revision period before the three observations of the last patient.

^b Items 19 and 32 added after first period of observation.

TABLE 1—(Continued)

Patient's major defenses shown were:

35. Projection (Attribution to other person of motives unacceptable to oneself)
36. Repression (Rationalization: logical excuse or justification of feelings or behavior)
37. Avoidance (Denial: avowed nonperception of reality situation, internal or external)
38. Intellectualization (Explanation of one's feelings or behavior in terms of general or theoretical principles or abstract concepts)
39. Isolation (Separation of feeling and idea)
40. Reaction formation (Turning into the opposite, internal)
41. Conversion (Displacement: shift of feeling from one object or person to another)
42. Patient will complete therapy.
43. Patient will need long-term therapy, 1½ years or more.
44. Diagnostic impression

Each of the four sections, then, consisted of three weekly observations of the same patient and therapist. After each triad of observations the data were examined and definitions were clarified with the goal of establishing a clearer basis for judgment. Between B_{1-3} and C_{1-3} most items were defined as a result of consultation with three psychoanalysts. New definitions were mimeographed on the rating sheets. Some of the less clear items were replaced by more nearly observational items.

The degree of interrater agreement was described in terms of the binominal expansion in which $p = q = .5$ for each item (present or absent), and n is the number of raters, generally six, but occasionally five. There is reason to question the .5 probability value for many clinical data since such data are apt to occur in a clinical sample with varying frequencies and clinicians are likely to have differential sets to perceive them. It seemed most reasonable, however, to treat the items as pennies, in the absence of other information.² Agreement among raters, then, was expressed at first in terms of p values obtained from the binominal expansion. We also wished to know whether our judges agreed more on the observational than on the inferential items. Hence compounding of probabilities was necessary. To compound the probabilities over several series, p values were converted into chi square according to Pearson's transformation: $\chi^2 = -2 \log_e p$ as described by Jones and Fiske (1953).³

² An empirical investigation of p and q was made by computing the proportion of present responses for each item over the 12 replications and setting up a distribution. The median value of .48 suggests that for the sample of items the assumption that $p = q = .5$ is a reasonable one.

³ The issue of independence of the tests is of importance here. Most of the items had been planned to be as probably independent as possible. Measurement of independence in this study is impossible because any item about which there is complete agreement will necessarily correlate perfectly, positively or negatively, with every other unanimous item. Whether the replications on the same patient can be considered independent events is still an open ques-

tion. Each chi square that enters into the compounding carries 2 *df*. Hence, the compound probability for a given item for three replications is the sum of the three chi square values and $df = 2 \times 3 = 6$. In similar fashion the compound probabilities of the amount of interrater agreement on the observational and inferential items were computed by totaling the chi square values for all observational and inferential items separately with *df* equal to twice the number of items.

The important questions were: do the judges agree more or less in their ratings of observational and inferential items, and do judges agree more in successive periods of observation as a function of experience and redefinition of terms? Statistically, these questions reduce to the significance of the difference between total amounts of agreement. Since our measures of agreement are expressed in terms of chi square, the statistical test is of the difference between two chi squares. To our knowledge, the only possible way of testing this difference is by means of the *F* ratio. The *F* ratio is defined (Peters & Van Voorhis, 1940, p. 420) as the ratio of two independent chi squares, each divided by its own *df*:

$$F = \frac{\chi^2_1/df_1}{\chi^2_2/df_2}$$

In this case χ^2_1 is the summation of the chi squares representing the amounts of agreement on all items in one treatment (e.g., observational items) and df_1 is the degree of freedom (e.g., twice the number of observational items). χ^2_2 and df_2 are the corresponding values for the other treatment (e.g., inferential items).

RESULTS

To attain a statistically significant degree of rater agreement in a given replication of a single item all raters must give the same judgment (for the values of n in this study). Unanimity yields a p value of .016 for six

tion. The nature of our findings, however, suggests that they were nearly independent.

TABLE 2
AMOUNT OF AGREEMENT AMONG JUDGES FOR EACH ITEM, PATIENT,
AND OBSERVATIONAL PERIOD

Item	Patient A				Patient B								Patient C				Total
	1	2	3	1-3	1	2	3	1-3	4	5	6	4-6	1	2	3	1-3	
Observational																	
1	.11		.11		.02	.02		.01	.02		.02	.01		.11			.001
6	.02		.02	.01									^a				
6a												^b		.02		.10	
6b												^b	.02	.02	.02	.001	
7			.02	.10	.02	.02	.03	.001	.02					.11	.11	.10	.001
9													^c				
9a												^b	.02	.02		.01	
9b												^b	.11	.11		.10	
9c												^b	.02	.11	.11	.01	
9d												^b	.02	.02	.02	.001	
9e												^b	.02	.02	.11	.01	
10	.02	.03	.02	.001	.11	.11			.11		.11	.10	.11	.02	.02	.01	.001
11		.03		.10	.02	.02	.03	.001	.02	.03	.02	.001					.001
12	.02	.03		.01					.11	.03		.10	.02	.11	.02	.01	.001
13	.02		.02	.01	.02	.02		.01		.03	.02	.01	.02	.02	.02	.001	.001
14		.03	.11	.05	.02	.02	.03	.001					.02	.11		.05	.001
15	.02	.03	.02	.001					.11		.11	.10	.02		.11	.05	.001
16					.11	.11		.10						.11			
17					.02	.11	.03	.01	.02		.02	.01	.11	.11	.02	.01	.001
18	.11		.02	.02	.02	.02	.03	.001	.02	.03	.02	.001		.02	.11	.05	.001
19	^b	.03	.02	.01		.02		.05	.02	.03	.02	.001	.02	.02	.02	.001	.001
20					.11	.11							.02	.11		.05	
21	.11				.02				.11	.03		.05	.02	.02	.11	.01	.001
32	^b	.03	.11	.05	.11		.03	.05			.02	.10	.11		.11	.10	.01
Inferential																	
2					.02	.02		.01	.02	.03	.02	.001			.02		.001
3-5	.02	.03	.02	.001	.02	.02	.03	.001	.02	.03	.11	.01		.11			.001
9	.11	.03	.11	.02		.11			.11		.11	.10	^a				
22		.03	.11	.05	.02			.10						.11		.10	.05
23	.02	.03	.02	.001	.11	.02		.05		.03			.02	.02	.02	.001	.001
24		.03	.11	.05	.11				.11		.02	.02	.11	.02	.02	.01	.001
25	.11		.02	.02	.11				.11	.03		.05	.11				.01
26	.02		.02	.01	.02			.05			.11			.02	.11	.05	.001
27	.02		.02	.01	.02			.05	.11						.11	.10	.001
28			.02	.05	.02		.03	.001		.03		.10	.02	.02	.02	.001	.001
29	.11	.03	.02	.01		.11							.11	.02	.02	.01	.001
30	.11				.02	.03	.01	.02	.03	.02	.001		.11				.001
31	.02	.03	.02	.001	.11				.11		.11		.02	.02	.02	.001	.001
33												^b		.11	.02	.05	
34a													.11	.11		.10	
34b	.02	.03	.11	.01	.02	.02		.01	.02	.03	.02	.001		.11	.11	.10	.01
35	.02	.03		.01	.02			.10	.11	.03		.05		.11			.01
36			.11								.11	^c		.11	.11	.10	
37												.05 ^c	.11		.11		
38		.03			.02	.02		.01	.02		.11		.02	.11		.05	.01
39	.11		.11	.10		.11	.03	.10					.02	.11		.01	.01
40	.02	.03	.02	.001	.02	.02	.03	.001	.02	.03	.02	.001	.02	.02	.11	.01	.001
41	.02		.02	.01		.11	.03	.05				^c			.02	.001	
42	.02	.03	.02	.001					.11				.11	.11	.11	.05	.001
43	.02	.03	.02	.001	.02	.02	.03	.001	.02	.03	.02	.001	.02	.02	.02	.001	.001

Note.—All entries are expressed in probability values.

^a Item deleted at this point.

^b Item added after this point.

^c Item replaced at this point.

raters and .031 for five raters. In only one of the 12 replications did more than half of the observational items attain this degree of rater agreement. The same is true of the inferential items. The proportion of observational items that showed significant agreement varied from 31.3% to 54.5% among the 12 observational periods. For the inferential items the range is from 28.0% to 54.2% (Table 2). It is patent that significant agreement was not the rule. On two items only was agreement unanimous throughout the 12 replications: absence of reaction formation (Item 40), and need for long-term therapy (Item 43).

Most, but not all, of the items were significantly in agreement when the probabilities were compounded over the 12 replications. Even so, most items varied enormously in degree of agreement from replication to replication; hence overall significance of agreement gives little ground for confidence at any one time or for any one case in clinical observation. Description of the vicissitudes of a few items may be informative. The presence of monotonous speech (Item 10) was significant in only 5 of the 12 replications, restlessness (Item 11) in 7 replications. Some content items were rarely significant. The problem area, hostility, a clinical favorite, was not once agreed upon unanimously. Presence of gestures (Item 20) was significantly agreed upon once in the series.

Among the inferential items depth of material (Item 22) was significant twice—before the clarifying definition. Presence of anxiety showed perfect agreement three times—early in the series; capacity for insight showed perfect agreement consistently for the first patient's three replications, and for the last patient's as well, and not at all for the second patient's six replications. Judgments about positive transference were never in complete accord; negative transference ratings were in accord for seven replications (all by ab-sence). Agreements in regard to ego defenses were as follows: reaction formation—perfect score, always absent; projection—four significant agreements split between two patients; denial, intellectualization, and isolation—each three times; rationalization—never significantly agreed upon.

While psychological defenses, it may be argued, are rather subtle and may become apparent only in intensive, therapeutically oriented observation, the lack of agreement in six successive observations periods with the same patient seems cause for some concern. The judges agreed only once in nine replications on the item: The hour was characterized by resistance. This item was replaced for patient C's observations by Items 9a through 9e which were deemed to represent some of the observational components of resistance. During the three observations of the last patient 9c was significant all three times, 9a and 9d twice, 9e once, and 9b not at all. If this finding can be generalized, it suggests that agreement about some clinical observations can be improved by specifying concrete behaviors.

Results of the comparison between observational and inferential items were unexpected. First of all, neither observational nor inferential items showed a significant preponderance of items with unanimous agreement as tested by a four-fold chi square test. When the combined probabilities of observational items were tested against those of the inferential items, not a single F ratio reached a .05 level of significance for the 12 replications individually, for any of the patient-therapist combinations, or for the 12 replications. Within the limitations of this experiment, then, there was no difference in rater agreement as a function of the degree of a priori objectivity of the items (Table 3).

There was, similarly, no significant difference in overall agreement (observational and inferential items combined) between patient-therapist combinations. That is, variations in the persons observed had no systematic effect upon the degree of agreement among the raters. Possible interactions with particular items may exist but are difficult to prove. And, finally and somewhat sadly, there was no improvement in degree of agreement as a result of practice, communication of criteria for rating each item, or specific definition of items. Some items improved and some deteriorated. In fact, the highest summated chi square for blocks of inferential or observational items or combinations of the two is

TABLE 3

SUMMATED χ^2 VALUES AND VARIANCES OF INTERJUDGE AGREEMENT IN RATINGS OF OBSERVATIONAL AND INFERENTIAL ITEMS

Observation Period	Observational			Inferential			Total		
	χ^2	V	df	χ^2	V	df	χ^2	V	df
A ₁	67.4	2.41	28	124.94	2.61	48	192.34	2.53	76
A ₂	62.81	1.96	32	115.95	2.42	48	178.76	2.23	80
A ₃	80.06	2.50	32	138.53	2.89	48	218.59	2.73	80
A ₁ -A ₃	210.26	2.29	92	379.42	2.63	144	589.69	2.50	236
B ₁	92.65	2.90	32	110.64	2.21	48	203.30	2.48	80
B ₂	82.33	2.57	32	107.34	2.15	48	189.68	2.31	80
B ₃	65.84	2.06	32	93.78	1.88	48	159.63	1.95	80
B ₁ -B ₃	240.83	2.51	96	311.77	2.08	144	552.60	2.25	240
B ₄	77.50	2.42	32	114.22	2.28	48	191.72	2.34	80
B ₅	62.25	1.95	32	104.77	2.10	48	167.02	2.03	80
B ₆	79.07	2.47	32	107.40	2.15	48	186.47	2.27	80
B ₃ -B ₆	218.82	2.28	96	326.39	2.18	144	545.20	2.22	240
B ₁ -B ₆	459.64	2.39	192	638.16	2.13	288	1,097.80	2.23	480
C ₁	125.82	2.86	44	109.52	2.28	48	235.34	2.56	92
C ₂	126.25	2.87	44	121.79	2.54	48	248.04	2.70	92
C ₃	103.28	2.35	44	119.82	2.5	48	223.1	2.43	92
C ₁ -C ₃	355.35	2.69	132	351.13	2.44	144	706.49	2.56	276
Total									
(A-C)	1,025.26	2.47	416	1,368.72	2.33	576			

not significantly different from the lowest value.

When probabilities are compounded over a series of replications, it is possible for much variation in agreement to occur among replications and still attain the .01 or .001 level of significance. To get the flavor of this variation it might be worthwhile to examine the behavior of one item. Item 31, capable of insight, was unanimously agreed upon for the three replications of Patient A; for Patient B the agreement was 5/6, 3/6, 4/5, 5/6, 3/5, and 5/6 (none of them significant); for Patient C all three replications were in total agreement. The compounded p value is beyond .001. Agreement was perfect for two patients and clearly in the direction of agreement, though not significantly, for the third patient. It may be argued that the nature of the patient is an important consideration. Perhaps so. Evaluation of insight possibilities in psychotic patients such as B may be less certain than in neurotics and the role of insight in improvement may be thought to differ as well. On Item 42 only the first three

replications yielded significant agreement, yet p reached .01.

In order to obtain some estimate of inter-rater agreement in quantitative terms, the judgments on all items for each judge were set up in four-fold tables singly with each other judge. Phi coefficients were computed for each pair of judges on each of the last three observation periods.

Phi coefficients ranged from .27 to .70 with median phi's for the three successive periods of .49, .45, and .40 in that order. Even though the phi coefficients are lower than Pearson's r 's would be, their values are higher than those of Gelfand, Quarrington, Widemann, and Brown (1954) on rating scales of Rorschach traits and Lisansky's (1956) phi coefficients for questionnaire items derived from the Rorschach. They also exceed the intercorrelations obtained by Stern, Stein, and Bloom (1956, p. 113) from Q sorts on the basis of school records, projective data, and behavioral observations. While 38 of our 45 phi's are significant at or beyond the .01

level, the magnitudes are still distressingly low for purposes of prediction.

One judge was consistently highest in his median ϕ 's with other judges. There was no consistency as to who correlated lowest with the others.

Of the items which were retained throughout the investigation, only two observational items (16 and 20) and one inferential item (34a) failed to be agreed upon beyond chance expectations. This finding can be interpreted in contradictory ways. Since the judges agreed beyond chance on most of the items over the whole experiment, they were evidently perceiving something in the way of communality. On the other hand, they were less consistently in agreement than seems desirable and their agreement fluctuated in no predictable fashion.

One can become lost in trivia and post hoc rationalizations in examining the behavior of specific items. Since a number of items showed unanimity in the judgments of the raters for one or more patients and not for others, it might be suspected that some patients present more clear-cut evidence about some clinical variables than other patients do and that patients differ in the kinds of clinical data for which they show evidence. To be sure, patients vary somewhat from hour to hour and it may be expected that their observers and therapists do also. Our data can be interpreted in whichever direction the reader's bias lies. It can be argued that the variations in amount of agreement from interview to interview render the clinical data practically useless for a given clinical situation. On the other hand, the fairly high level of agreement over the 12 replications indicates that there is significant communality of clinical perception.

It should be remembered that the maximum number of raters was six and that the divergent opinion of one rater makes quite a difference in the amount of agreement. A larger number of raters would lessen the effect of a single rater.

DISCUSSION

A rough generalization that can be made from these findings is that for *most* of the items the judges agreed in the combined

data beyond chance expectations, but that the degree of agreement varied from item to item, patient to patient, and replication to replication in no predictable fashion as others have found (Forer, Farberow, Meyer, & Tolman, 1952; Gelfand et al., 1954) in their study of Rorschach ratings. Such unsystematic variation in agreement does not necessarily mean that clinical observation is too subjective to be of practical significance. It does mean, however, that some of the parameters of clinical observation and inference could, perhaps, be profitably re-examined and reformulated.

We might ask ourselves whether the present experiment is a fair test of the clinical interview. Was the situation real enough; did it tap variables that are ordinarily involved in therapists' observations? Would a therapist ask himself such questions during a therapy session, or are these judgments generally the result of summarizing observations gleaned from long-term contact with the patient? Yet, many of these variables are assessed at the end of single initial interviews with reference to diagnostic or therapeutic goals.

Studies of the accuracy of clinical judgment and prediction have yielded little evidence of general superiority attributable to professional training (Cline, 1955; Grigg, 1958; King, Ehrmann, & Johnson, 1952; Lisansky, 1956; Luft, 1950, 1951). Our judges' senior status suggests a fairly high level of professional competence, but it does not imply that further development of skills or radical changes in orientation are unlikely to occur, even though we would ordinarily assume that they had reached an asymptote in their perception of most of the variables used in this experiment. Evidence is not impressive that homogeneity of training necessarily conduces to homogeneity of judgment. Is it, then, a fact that for certain kinds of clinical judgment the variance attributable to individual differences among judges exceeds that attributable to professional training? It seems so in this study. We are forced to agree with Bendig (1956) that individual differences among judges may outweigh many facets of the rating process and the data to be judged.

Discussion of terms after each series of three observation periods had no measurable

effect on degree of agreement. Lisansky (1956) believes that improvement can occur in the rating of Rorschach protocols, despite her and our (Forer et al., 1952) empirical findings to the contrary. Wiener's (1958) belief that "We can and must train ourselves to agree on the judgments we make from projective test protocols" seems more a wish than a likely prospect.

There is a possibility that amount of rater agreement is inversely related to the amount of data which the clinician must process. The task of sorting a large mass of heterogeneous data may create interference with the evaluation of any one of the classification variables. Evidence suggests that there may have been too much rather than too little information, possibly of a contradictory nature, and too many ratings competing for the observer's attention (Borke & Fiske, 1957; Cutler, Bordin, Williams, & Rigler, 1958; Gage, 1953; Giedt, 1955; King et al., 1952; Kostlan, 1954; Luft, 1950, 1951).

Perhaps clinicians need to take stock of what they are asking from themselves, to appraise realistically rather than hopefully what is possible, so that they need not be unduly apologetic nor defensively nihilistic toward research evidence that questions their prowess.

It seems unlikely from this and other studies that any conceptual system or any amount of training can engender the degree of conformity or reproducibility in clinical perception and judgment that is achieved by standardized tests or electronic computers. Would such a state of affairs be desirable? The growing supplementation and frequent replacement of objective tests by projective methods in diagnostic work suggests that objective methods leave something to be desired. The price of objectivity is limitation of information, and the clinician feels the need for more and different kinds of information than that provided by objective tests, even though the information be of a lower order of reliability. Factually he deals with patients' verbalizations which are also of a low order of reliability and it is through his inferences that the clinician constructs a relatively stable conceptual model of his patients.

Complete unanimity of clinical judgment would represent constriction of the range of cues and of clinical attention, hence of therapeutic activity. Zero variance among therapists would likely generate low variance in therapeutic activity. But therapists differ inevitably as persons, in their preferred theoretical systems, in their ability to use particular techniques, and in their apparent effectiveness in dealing with different kinds of patients. The all-around therapist who works equally well with all kinds of patients is as much a myth as the psychological test that measures every aspect of the psyche.

It may be that the less than perfect agreement in the clinical observations described above reflects those individual differences among therapists that enable them to specialize, learn from one another, grow continually in their skills, and discover new concepts and techniques.

SUMMARY

As a means of investigating the reliability of psychologists' perception of clinical data, six diplomates in clinical psychology observed three patient-therapist teams for a total of 12 weekly psychotherapy sessions. Independent ratings of present or absent were made on a check-list containing a number of presumably observational and inferential items. After each series of three sessions the clinicians discussed, redefined, and replaced items with the goal of increasing interjudge agreement.

1. On very few items was there consistently significant agreement among the judges.
2. The amount of agreement on most items varied from session to session and patient to patient in no detectable pattern.
3. While the amount of agreement compounded over the 12 sessions was significantly beyond chance expectations for most items, it was not sufficiently substantial to warrant confidence in the judges' observations.
4. The amount of agreement among judges was not affected by the apparent objectivity of the item. Even more precisely operationally-defined items—such as silence of 30 seconds or more—were not consistently agreed upon.

5. There is no evidence that practice in judging, increased contacts with a particular patient, or discussion and clarification of items enhance objectivity.

REFERENCES

- BENDIG, A. W. The personality of judges and their agreement with experts in judging clinical case histories. *J. consult. Psychol.*, 1956, 20, 422.
- BORKE, HELENE, & FISKE, D. W. Factors influencing the prediction of behavior from a diagnostic interview. *J. consult. Psychol.*, 1957, 21, 78-80.
- CLINE, V. B. Ability to judge personality assessed with a stress interview and sound film technique. *J. abnorm. soc. Psychol.*, 1955, 50, 183-187.
- CUTLER, R. L., BORDIN, E. S., WILLIAMS, JOAN, & RIGLER, D. Psychoanalysts as expert observers of the therapy process. *J. consult. Psychol.*, 1958, 22, 335-340.
- FORER, B. R. Psychological test reports: Universal or discriminating. *J. nerv. ment. Disease*, 1959, 129, 83-86.
- FORER, B. R., FARBEROW, N. L., MEYER, M. M., & TOLMAN, RUTH S. Consistency and agreement in the judgment of Rorschach signs. *J. proj. Tech.*, 1952, 16, 346-351.
- GAGE, W. L. Explorations in the understanding of others. *Educ. psychol. Measmt.*, 1953, 13, 14-26.
- GELFAND, L., QUARRINGTON, B., WIDEMANN, H., & BROWN, J. Interjudge agreement on traits rated from the Rorschach. *J. consult. Psychol.*, 1954, 18, 471.
- GIEDT, F. H. Comparison of visual, content, and auditory cues in interviewing. *J. consult. Psychol.*, 1955, 19, 407-416.
- GRIGG, A. E. Experience of clinicians, and speech characteristics and statements of clients as variables in clinical judgment. *J. consult. Psychol.*, 1958, 22, 315-319.
- JONES, L. V., & FISKE, D. W. Models for testing the significance of combined results. *Psychol. Bull.*, 1953, 50, 375-382.
- KING, G. F., EHRLMANN, J. C., & JOHNSON, D. M. Experimental analysis of the reliability of observations of social behavior. *J. soc. Psychol.*, 1952, 35, 151-160.
- KOSTLAN, A. A method for the empirical study of psychodiagnosis. *J. consult. Psychol.*, 1954, 18, 83-88.
- LISANSKY, EDITH S. The inter-examiner reliability of the Rorschach test. *J. proj. Tech.*, 1956, 20, 310-317.
- LUFT, J. Implicit hypotheses and clinical predictions. *J. abnorm. soc. Psychol.*, 1950, 45, 256-259.
- LUFT, J. Differences in prediction based on hearing versus reading clinical interviews. *J. consult. Psychol.*, 1951, 15, 115-119.
- PETERS, C. C., & VAN VOORHIS, W. R. *Statistical procedures and their mathematical bases*. New York: McGraw-Hill, 1940.
- STERN, G. G., STEIN, M. I., & BLOOM, B. *Methods in personality assessment*. Chicago: Free Press, 1956.
- WIENER, M. The problem of interjudge agreement and prediction. *J. proj. Tech.*, 1958, 22, 122-123.

(Received April 11, 1960)

AUTHORITARIANISM IN THE THERAPEUTIC RELATIONSHIP¹

JOHN L. VOGEL

University of Washington

This study was concerned with the therapist, the patient, and their relationship in psychotherapy. It dealt with authoritarianism as a personality trait in each of the individuals, and tested for associations between authoritarianism as a trait, attitude, and behavior. A major hypothesis of this study was that the peculiar interaction of authoritarianism in therapist and patient would be crucial to the development of the therapeutic relationship. Although this study did concern itself with authoritarianism, this trait was not necessarily thought to be the most basic or critical aspect of the therapeutic relationship. It was selected for study here to demonstrate the importance of considering the personality, needs, and motives of therapist and patient, as they interact in the therapeutic relationship.

Of the many personality variables that might be studied in this manner, the writer chose to consider authoritarianism, as delineated in the major work on *The Authoritarian Personality* (Adorno, Frenkel-Brunswik, Levinson, & Sanford, 1950). It was thought that the patient population of any clinic might not be as individualistic, equalitarian, and self-actualizing as some writers seemed to assume. Further, even a generally equalitarian patient may develop authoritarian expectations about psychotherapy from his experience with other professions. For the therapist, we know that there is a wide range of therapeutic behavior in terms of training and orientation, to say nothing of the range of attitudes and needs they may have. Authoritarianism, then, was thought to be one of the

trait dimensions relevant to patient and therapist roles.

A major issue which still surrounds authoritarianism as measured by the F Scale refers to the question of its behavioral implications and correlates. Titus and Hollander (1957) raise serious question about the relationship between authoritarian attitudes and behavior. They urge special caution where interpersonal behavioral implications are to be drawn. Christie, on the other hand, makes a strong case for congruency of F Scale scores and predicted behavior, citing four studies in support of his position (Christie & Jahoda, 1954, p. 145). As a test of the question, this study hypothesized that authoritarianism, as a trait of therapists and patients, would find expression in their attitudes toward psychotherapy and their behavior in therapy.

The second basic hypothesis of this study follows from the argument that a similarity of personality traits in patient and therapist tends to facilitate the relationship. Barron's thesis (1950) seems to be the first study to consider both patient and therapist variables in an experimental approach to the therapeutic relationship. Axelrod (1951) argued and found partial support for the hypothesis that progress in therapy was more likely when the personalities of patients and therapists were similar than when they were dissimilar. Underlying this hypothesis was the theory that

the presence of an emotional identification or empathy between patient and therapist, springing from common emotional experience and manifested more or less by a similarity of personalities, is a condition favorable for the successful development of the therapeutic process (pp. 4-5).

Studies by Bown (1954), Hiler (1958), Libo (1957), and Ashby, Ford, Guernsey, and

¹ Based on a doctoral dissertation submitted to the University of Chicago, 1959. The writer is indebted to Donald W. Fiske, Desmond S. Cartwright, and Ralph W. Heine for their encouragement and help.

Guerney (1957) are pertinent considerations of this question. Although the evidence is something less than substantial, there does seem to be a line of thought suggesting that there is an interaction between the personality traits of therapist and patient, and that generally a similarity of traits tends to facilitate the relationship. This position receives some support from studies in the fields of leadership and education (Goldberg & Stern, 1952; Haythorn, Haefner, Langham, Couch, & Carter, 1956; Jones, 1954; Sanford, 1950). The second basic hypothesis of this study states that a similarity of therapist and patient along the trait dimension of authoritarianism-equalitarianism is related to the establishment of successful or good therapeutic relationships.

Sanford (1956) raises some question about whether authoritarian patients, without reference to therapist traits, may not have real difficulty forming therapeutic relations with any therapist. He is rather blunt on this point, writing:

The person high on F rarely seeks, but rather resists the idea of psychotherapy; and once a start has been made, the technical problems are trying (p. 313).

Sanford then goes on to note a study by Freeman and Sweet (1954) in which they offered evidence that patients with many features of the F pattern actually respond better in certain forms of group therapy than they do in individual therapy. This argument, obviously, refers to patient traits only. As a more parsimonious explanation of therapeutic failure the question merits testing and forms a specific hypothesis of this study.

METHOD

Instruments

The California F Scale as a measure of authoritarianism was taken directly from Forms 45 and 40 as published by Adorno et al. (1950). One item: "It is best to use some prewar authorities in Germany to keep order and prevent chaos," was omitted as untimely and probably ambiguous to most subjects. Scores were derived in the conventional manner. Responses from -3 to +3 were converted to positive scores ranging from 1 to 7, with no response scored as 4. The sum of scores for the 29 items was used for tests of hypotheses in this study.

The Authoritarian-Equalitarian Therapy sort (designated as AET) was especially developed for this

study. A 40-item card sort was constructed containing 20 items reliably prejudged as descriptive of an authoritarian therapy relationship, and 20 items similarly prejudged as descriptive of an equalitarian therapy relationship. By verbal instruction the subjects were asked to sort the 40 items in 8 piles of 5 items each. The piles were numbered from 1 to 8, pile Number 1 designated as "Least True or False," pile Number 8 as "Most True." Patients were asked to sort the 40 items to indicate "which of these things you would like to have be most true and which of these things you would like to have be least true, or even false, about the relationship between you and your doctor (therapist, counselor)." Therapists were given essentially similar instructions with added emphasis on the expression of "own opinions" rather than what they had been taught or had read. Each item was scored according to the pile number in which it was placed. The scores for the 20 authoritarian items were summed for each subject and designated the AET score, with a possible range from 50 to 130. For each patient-therapist pair an AET Discrepancy Score was computed by summing the squares of the score differences over the 40 items. This Discrepancy Score is, of course, a negative function of the correlation between patient and therapist sorts. The Discrepancy Score was considered an adequate representation of similarity and differences of patient and therapist attitudes toward therapy along the specific dimension of authoritarian-equalitarian attitudes and behaviors.

A Therapist Rating Scale was developed, drawing heavily from an instrument developed at the University of Chicago Counseling Center (Rogers & Dymond, 1954, p. 101) and currently in use there. Several items of the original form were omitted to produce a shorter rating blank. A new item was introduced in which the therapist was asked to rate the "quality of the relationship," thus: "Does this seem to be a 'good' and effective therapeutic relationship? How do you estimate the quality of the therapeutic relationship between yourself and this patient?" (nine-point scale from "poor" to "good"). The rater's estimate of patient satisfaction in the relationship was retained in its original form, thus: "Estimate the patient's feeling about the relationship" (nine-point scale from "strongly dissatisfied" to "extremely satisfied"). Only these two items are utilized in the present study.

The formation of successful or better therapeutic relationships as a criterion was assumed to be directly related to the various types of criteria employed in other studies, but it was thought to have a specific pertinence of its own, as elemental or more basic. It seemed reasonable to attempt a direct measure of the quality of the relationship. It was assumed that the quality of the relationship is largely determined and may be evaluated in the very early contacts of patient and therapist. Although a rating of patient satisfaction may not be one of the essential goals of psychotherapy and may not be directly related to the quality of the relationship, it was thought to be a useful supplementary criterion meas-

ure. No matter how good the quality of the relationship may appear to the therapist or judges, the degree of patient satisfaction with its implications for continuance in therapy or premature termination may be a crucial evaluation.

An Observer Rating Scale was developed for the use of judges in rating patient and therapist behaviors as observed on short recorded segments of therapy. Items 1 and 2 provided estimates of the quality of the relationship and patient satisfaction, and were identical in form to the items described above. In Item 3 the therapist's behavior in the recorded segment was rated on five dimensions: aggressive-submissive, directive-nondirective, highly anxious-low anxiety, dominating-equalitarian, and rigid-flexible. In Item 4 the patient's behavior was rated on these five dimensions: aggressive-submissive, dependent-self-sufficient, highly anxious-low anxiety, conventional-individualistic, and rigid-flexible. From the many qualities and behaviors attributed to authoritarians in the literature, these five in each case were selected as being both relevant and ratable. In Item 5 the judge was asked to rate the behavior of the therapist along the single dimension of authoritarian-equalitarian on a nine-point scale. In Item 6 a distinction was made between dominant and submissive types of authoritarian behavior by the patient. Dominant behavior was defined by aggressive active authoritarian behavior, while submissive behavior was defined by passivity or deference, expecting or seeking authoritarian behavior by the other. Although dominant and submissive authoritarian behaviors were thought to be dynamically related, it seemed plausible to consider the two aspects mutually exclusive in any short sample of behavior. Thus, a V-shaped scale was used, with equalitarian at the apex and authoritarian-dominant and authoritarian-submissive at each of the two extensions, each on a nine-point scale. The judge was asked to select the aspect most prominent in the given segment and make a rating on the selected scale.

Samples

The subjects were drawn from two clinic populations. Those designated as Group A include treatment cases in the Psychiatry Clinic of Albert Merritt Billings Hospital, University of Chicago. Senior medical students are required, as part of their training in psychiatry, to treat in psychotherapy one selected patient who has been referred to the clinic. It should be noted that these students had little training and no prior experience in psychotherapy. All patients were told that their treatment would be limited to 18 weeks' duration, after which they would be either terminated or referred elsewhere. The present sample is composed of patients and therapists drawn from this program during two successive quarters. Of the 35 patients originally tested for this study, 1 was eliminated because of a suggested organic involvement, 1 for an alleged inability to read, and 1 patient who failed to keep the first and subsequent therapy appointments. The remaining sample of 32

cases included 15 males and 17 females, with a mean age of 38 years, ranging from 23 to 68 years.

The subjects designated as Group B were drawn from the client population of the University of Chicago Counseling Center. Clients are normally assigned to therapists on the basis of therapist availability, and clients who agree to participate in research studies are then randomly assigned to projects in progress at that time. The present sample includes 30 cases assigned to the writer's project over a 6-month period. The therapists in this group included three staff members with extensive experience, seven staff members with some or considerable experience, and seven students in training who were seeing their first or second cases. The client population included 16 males and 14 females, with a mean age of 27 years, ranging from 19 to 43 years.

The population in Group A includes 32 patients and 32 therapists, each patient seeing a different therapist. In Group B, the population includes 30 clients and 17 therapists, several therapists treating more than one client in this sample.

Collection of Data

Patients and therapists were seen prior to their first therapeutic interview and were asked to complete the F Scale and AET sort. After the second therapeutic interview, the therapist completed the Therapist Rating Scale.

Observer ratings were made on Group A only. Recordings of the first interview were retained, and 5-minute segments were selected from the beginning and ending of each interview. These segments were rerecorded in random order, with at least five other segments between the two segments of any given interview. Two judges (the writer and another graduate student of psychology, both with training and experience in psychotherapy) rated each segment on the Observer Rating Scale. Thus, for each case there were four ratings: beginning and ending segments by each of two judges. One recording was inaudible and tests based on judges' ratings will be drawn from an N of 31. Reliability of judges' ratings was tested on each of the 14 scales of the rating form. The two judges' summed ratings (beginning plus ending segments) were significantly correlated on 10 of the 14 (9 at the .01 level, 1 at the .05 level: $r = .38$). Only these 10 items were utilized in this study. It is striking that the four items on which the judges were not in agreement all dealt with patient traits.

RESULTS

Authoritarianism, as a personality trait of the therapist, was hypothesized to be significantly related to his description of the ideal therapeutic relationship in terms of directive, paternalistic, and nurturant qualities. Full scale scores on the F Scale were compared to AET scores. For the 32 therapists in Group A the Pearson r correlation was .03, a clearly

nonsignificant result. For Group B, with 17 therapists, the Pearson r correlation was .62, significant at the .01 level.

It was predicted that therapists characterized by authoritarianism would tend to show more authoritarian behavior in their therapy than those characterized as equalitarian. The Observer Rating Scales were utilized here. The 31 therapists were dichotomized on the basis of their F Scale scores, 16 low and 15 high. Results in tests of this hypothesis may be summarized as follows: (a) On a global behavioral rating of authoritarian-equalitarian the high F scorers were rated significantly more authoritarian than low F scorers. (b) Although high and low F scorers did not differ on the full scale dimension of aggressive-submissive, they did differ on their deviation from "appropriate" mid-point behavior, i.e., high scorers were given more extreme ratings on this dimension. (c) High scorers were rated as more directive, anxious, and were rated as more dominating than low scorers, but not significantly so. (d) Behavior of high scorers was rated as significantly more rigid than that of low scorers.

Authoritarianism, as a personality trait of the patient, was hypothesized to be significantly related to his description of the ideal or preferred therapeutic relationship in terms of directive, paternalistic, and nurturant qualities. Full scores on the F Scale were compared to AET scores. The 32 patients in Group A showed a Pearson r correlation of .34, significant at the .05 level. In Group B, with 30 patients, the Pearson r correlation was .38, significant at the .05 level.

It was predicted that patients characterized by authoritarianism would tend to show more authoritarian behavior in their therapy than those characterized as equalitarian. The Observer Rating Scales were utilized here: a global rating of patient behavior on a nine-point scale and a rating on patient aggression. As a test of this hypothesis the 31 cases were dichotomized on the basis of the patient's F Scale score, 15 low and 16 high. On the global rating of patient behavior the difference between low and high scorers was not significant. The two groups did not differ on the full scale dimension of aggressive-submissive. High scoring patients did show the larger

deviation from "appropriate" mid-point behavior as predicted, but the difference between groups was not significant.

In line with the argument of Sanford, discussed above, it was predicted that patients who are characterized as equalitarian will tend to form better therapeutic relationships than those characterized as authoritarian. In Group A, the hypothesis was tested against four criterion measures: the therapist's rating of the quality of the relationship, therapist's estimate of patient satisfaction, judges' composite rating of the quality of the relationship, and the judges' composite estimate of patient satisfaction. The differences between low and high scoring patients on the therapist ratings were not significant. The differences on judges' ratings were both in the predicted direction. Judges rated the quality of the relationship significantly ($t = 2.50$, $p < .01$) higher for the group of low F scorers, and the estimate of patient satisfaction was slightly higher for this group but not significantly so. In Group B the hypothesis was tested against two criterion measures, the therapist's rating of the relationship and his estimate of patient satisfaction. Differences between low and high scorers were not significant.

The last three hypotheses were developed from the argument that similarity of patient and therapist personalities facilitates the development of good therapeutic relationships. It was hypothesized that patients characterized by authoritarian traits would tend to form better therapeutic relationships with therapists characterized as authoritarian than with those characterized as equalitarian. For a test of this and the following hypothesis the dichotomies between high and low scorers in patient and therapist groups were retained. First, each of the patients characterized as authoritarian was considered with his respective therapist. Mean criterion ratings are shown in Table 1. In Group A the hypothesis was tested against the four criterion measures listed above. All differences were nonsignificant. In Group B the hypothesis was tested against the two criterion measures listed above. Both differences were nonsignificant.

Secondly, it was hypothesized that patients characterized by equalitarian traits would tend to form better therapeutic relationships

TABLE 1

MEAN CRITERION RATINGS ON AUTHORITARIAN AND EQUALITARIAN GROUPS OF PATIENTS WITH THEIR RESPECTIVE AUTHORITARIAN AND EQUALITARIAN THERAPISTS

	Criterion					
	Group A				Group B	
	Therapist Rating		Observer Rating		Therapist Rating	
	QR ^a	PS	QR	PS	QR	PS
Authoritarian Patients						
Authoritarian Therapist	5.89	6.11	3.44	4.95	5.80	6.20
Equalitarian Therapist	7.14	6.71	4.14	5.57	5.00	5.25
Equalitarian Patients						
Equalitarian Therapist	5.89	5.55	5.39	5.97	5.70	5.40
Authoritarian Therapist	6.00	6.14	4.38	4.75	6.28	6.86*

^a QR = Quality of Relationship, PS = Patient Satisfaction.
* Difference significant at .05 level, in a direction opposite to that predicted.

with therapists characterized as equalitarian than with those characterized as authoritarian. Each of the patients characterized as equalitarian was considered with his respective therapist. Mean criterion ratings are shown in Table 1. For Group A, on the four criterion measures, all differences were nonsignificant. In Group B both therapist ratings were in a direction opposite to that predicted, with the difference on rated patient satisfaction significant at the .05 level.

In the last hypothesis, therapist and patient descriptions of ideal or preferred therapy conditions (AET) were utilized. Discrepancy Scores for each case were computed as

previously described. These scores were dichotomized in terms of low and high discrepancy. It was predicted that the quality of the therapeutic relationship would be related to the degree of discrepancy between patient and therapist expectations of authoritarian attitudes and behavior in therapy. Mean criterion ratings of low and high discrepancy groups are shown in Table 2. Although only one of the differences was statistically significant, low discrepancy cases received higher ratings on all criterion measures in both groups.

DISCUSSION

The failure to find a relationship between F Scale scores and attitudes toward therapy in the therapist population of Group A may be related to the nature of the F Scale items and the students' reaction to them. It has been said

TABLE 2

MEAN CRITERION RATINGS ON CASES WITH HIGH AND LOW DISCREPANCY BETWEEN THERAPIST AND PATIENT AET SORTS

Patient Group	Criterion					
	Group A				Group B	
	Therapist Rating		Observer Rating		Therapist Rating	
	QR ^a	PS	QR	PS	QR	PS
High Discrepancy	6.06	6.00	4.16	5.12	5.00	5.53
Low Discrepancy	6.31	6.19	4.55	5.58	6.33*	6.13

^a QR = Quality of Relationship, PS = Patient Satisfaction.
* Difference significant at .05 level.

that authoritarian people as measured by the scale agree more with authoritative statements; and that, therefore, a portion of the discriminatory power of the F scale derives from its form, rather than its content (Leavitt, Hax, & Roche, 1955, p. 221).

The very authoritative tone of the statements in the F Scale, referred to as a form characteristic, may, however, operate with reactive effect on some subjects. Several of the thera-

pists (who, it will be recalled, were senior medical students) commented on the stringent wording of the statements. One student commented that: "In medical school one of the first things you learn is to suspect any statement with 'always' or 'never' in it." These are not individuals who are rigidly or self-consciously equalitarian, but rather students trained to be critically sensitive to the literal meaning of words, and to hold in suspicion all authoritative sounding statements. The form component may, in such cases, have an inhibitory, and thus invalidating, effect.

Since therapists' scores on the F Scale do correlate quite well with their rated behavior in therapy it may be more reasonable to view their F Scale scores as a relatively reliable representation of authoritarianism as a personality trait and to re-evaluate their expression of attitudes toward therapy. It is well to remember that this population of therapists is composed of students with no experience and very little training in psychotherapy. They probably had few consciously developed attitudes toward therapy. By contrast, the therapists in Group B, with more training and experience in therapy, do show a consistency between personality trait and therapy attitudes. It may be proposed that one of the consequences of training and experience is the increased congruence of therapist traits and attitudes, a greater consistency between the personality of the therapist and his consciously held and expressed attitudes toward therapy. Whether such congruency is an effect of training or experience, or both, could and should be tested.

It was noted that the judges rated the quality of the relationship significantly higher for the patient group of low F scorers, while differences between therapist ratings were not significant. It may be that this reflects some differences in conception of the "good patient" role, and differences in what constitutes a "good and effective therapeutic relationship." Some differences in perspective between therapist and judges may also be operative here.

The finding that the rated quality of the relationship is related to the degree of similarity of patient and therapist descriptions of a preferred relationship on items specifically

defining authoritarianism tends to support the second basic hypothesis of this study. The quality of the relationship and an estimate of patient satisfaction in the early interviews appear to be somewhat predictable. To say this in another way: there does seem to be some pretherapy data from which we could anticipate good or poor, satisfying or unsatisfying, therapeutic relationships.

An observation may be made on the failure to find a relationship between the criterion and similarity on F Scale scores. Dichotomizing cases at the mean F Scale score for the group is probably too gross a division. For individuals not scoring in the extreme, high or low, authoritarianism is probably not the most crucial trait. The writer would speculate that for these individuals there are other traits, attitudes, and needs which play a more crucial role in determining the quality of their interpersonal relationships.

It may also be observed that attitude items, the AET sort, have a greater immediacy or relevance to the therapy situation than F Scale items. Many AET items refer to attitudes or behaviors which are very soon conspicuous by their presence or absence. By contrast, the F Scale measures a more fundamental trait which may not express itself so immediately or directly. In spite of the careful manner in which the AET items were developed, it may be that the sort contains several items of serious import to the development of the relationship, but not heavily loaded with authoritarianism. The method of deriving the Discrepancy Score, by summing the squares of the pile number differences over all items, gives an equal impact to all items.

This discussion should not, however, obscure the finding that similar attitudes of therapist and patient toward therapy were related to better therapy relationships. We are still some way from the point at which we can "match" patient and therapist to maximize success in therapy. As a therapist, the writer doubts that research of this kind will ever take all of the "mystery" and the essentially personal quality out of psychotherapy. Research may, however, help us to avoid the more blatant difficulties, and thus permit the more individual aspects of psychotherapy to operate more effectively.

SUMMARY

It was predicted that authoritarianism, as a personality trait of therapist and patient, would be reflected in their attitudes toward therapy and in their therapeutic behavior. Secondly, it was hypothesized that authoritarianism and equalitarianism, as interacting personality traits of therapist and patient, would have specified effects upon the quality of the relationship established.

A total of 62 patients and 49 therapists in two clinic populations completed the California F Scale and a specially devised instrument in which they described the ideal or preferred therapeutic relationship. After the second interview these therapists completed a scale containing two criterion items: a rating of the quality of the relationship and an estimate of patient satisfaction with the relationship. In one of the two clinic settings, two 5-minute segments were selected from each of the first interview recordings. For each segment, two judges rated the two criterion items and specific and general traits referring to authoritarian behavior on the part of the therapist and patient.

Authoritarianism (as measured by the F Scale) was found to be related to authoritarian attitudes toward therapy in both patient populations and in one of the two therapist populations. The hypothesis that authoritarianism, as measured by the F Scale, would be related to authoritarian behavior in therapy was supported for the therapist population, but not for the patients. A test of the hypothesis that equalitarian patients would form better therapeutic relationships than authoritarian patients gave equivocal results. The second basic hypothesis, that similarity of therapist and patient on the specific dimension of authoritarian-equalitarian would tend to facilitate the relationship, was not supported. There was, however, an association between criterion ratings and the amount of discrepancy between therapist and patient descriptions of the ideal or preferred relationship on items related to authoritarianism.

REFERENCES

- ADORNO, T. W., FRENKEL-BRUNSWIK, ELSE, LEVINSON, D. J., & SANFORD, R. N. *The authoritarian personality*. New York: Harper, 1950.
- ASHBY, J. D., FORD, D. H., GUERNEY, B. G., JR., & GUERNEY, LOUISE F. Effects on clients of a reflective and a leading type of psychotherapy. *Psychol. Monogr.*, 1957, 71(24, Whole No. 453).
- AXELROD, J. An evaluation of the effect on progress in psychotherapy of similarities and differences between the personality of patients and their therapists. Unpublished doctoral dissertation, New York University, 1951.
- BARRON, F. X. Psychotherapy as a special case of personal interaction: Prediction of its outcome. Unpublished doctoral dissertation, University of California, 1950.
- BOWN, O. H. An investigation of the therapeutic relationship in client-centered psychotherapy. Unpublished doctoral dissertation, University of Chicago, 1954.
- CHRISTIE, R., & JAHODA, M. (Eds.) *Studies in the scope and method of the authoritarian personality*. Glencoe, Ill.: Free Press, 1954.
- FREEMAN, M., & SWEET, B. A theoretical formulation of some features of group psychotherapy and its implications for selection of patients. *Int. J. group Psychother.*, 1954, 4, 355-368.
- GOLDBERG, S., & STERN, G. The authoritarian personality and education. *Amer. Psychologist*, 1952, 7, 375. (Abstract)
- HAYTHORN, W., HAEFNER, D., LANGHAM, P., COUCH, A., & CARTER, L. The effects of varying combinations of authoritarian and equalitarian leaders and followers. *J. abnorm. soc. Psychol.*, 1956, 53, 210-219.
- HILER, E. W. An analysis of patient-therapist compatibility. *J. consult. Psychol.*, 1958, 22, 341-347.
- JONES, E. E. Authoritarianism as a determinant of first-impression formation. *J. Pers.*, 1954, 23, 107-127.
- LEAVITT, H. J., HAX, H., & ROCHE, J. H. "Authoritarianism" and agreement with things authoritative. *J. Psychol.*, 1955, 40, 215-221.
- LIBO, L. M. The projective expression of patient-therapist attraction. *J. clin. Psychol.*, 1957, 13, 33-36.
- ROGERS, C. R., & DYMOND, ROSALIND F. (Eds.) *Psychotherapy and personality change*. Chicago: University of Chicago Press, 1954.
- SANFORD, F. H. *Authoritarianism and leadership*. Philadelphia: Institute for Research in Human Relations, 1950.
- SANFORD, N. The approach of the authoritarian personality. In J. L. McCary (Ed.), *Psychology of personality: Six modern approaches*. New York: Logos, 1956. Pp. 253-319.
- TITUS, H. E., & HOLLANDER, E. P. The California F Scale in psychological research: 1950-1955. *Psychol. Bull.*, 1957, 54, 47-64.

(Received February 5, 1960)

SOCIAL DESIRABILITY AND RESPONSE TO PERCEIVED SITUATIONAL DEMANDS

DAVID MARLOWE

AND

DOUGLAS P. CROWNE

College of Medicine, University of Kentucky

Ohio State University

Current research on social desirability (Cowen & Tongas, 1959; Edwards, 1957; Wiggins & Rumrill, 1959) has been chiefly concerned with a descriptive analysis of the influence of this variable on personality test responses. Along these lines, social desirability has achieved major status as a psychometric variable, the properties typically ascribed to it being those of a *stylistic* response determinant (Jackson & Messick, 1958). Pre-eminently, social desirability is considered to be a characteristic of *test items* (Edwards, 1957), and two models have been applied to its assessment. In the first of these procedures, items on a test are rated for social desirability by judges, and then responded to by subjects (Ss) under standard instructions (Edwards, 1953; Rosen, 1956); the correlation of the two sets of responses is inferred to indicate the amount of test response variance accounted for by social desirability. The second model involves the development of rational or empirical social desirability scales (Edwards, 1957; Hanley, 1957), the items of which show marked social desirability properties. Correlations between these scales and various personality tests, such as the MMPI, are assumed to reflect social desirability bias in the test responses. This method, however, can also be employed to identify dissimulators, i.e., those Ss whose personality test responses conform to the cultural stereotypes represented by the social desirability scale (Wiggins, 1959).

The prevalent conceptions of social desirability thus reflect an exclusive concern with response distortion in psychometric situations, with an attendant narrowing of research interests to investigations of the social desirability scalability of test items. The concept of social desirability has not been systemati-

cally investigated in terms of the *motivation* of Ss to dissimulate on personality tests and the relevance of this motivation to behavior in other, nontest situations.¹ This latter conception suggests research in which the differential influence of the need to respond in a socially desirable fashion would be investigated in situations where "self" or "item" evaluation is not the primary dependent variable. The present experiment was undertaken with this view in mind.

In a recent report, the writers (Crowne & Marlowe, 1960) described the development and preliminary validation of a new social desirability scale (M-C SDS) and outlined the construct of which the scale is at present the sole operational definition. In the initial study, however, only the essentials of a motivational concept of social desirability were suggested, and it is desirable here to present in further detail some of the implications of the construct.

Social desirability, as presently defined, refers to a need for social approval and acceptance and the belief that this can be attained by means of culturally acceptable and appropriate behaviors. In a psychometric situation, a high need for social approval would be inferred from a person's attribution

¹ Shortly after the completion of this experiment, Allison and Hunt (1959) reported a study investigating the relationship between Edwards SDS and aggressive responses to varying conditions of frustration as measured by a paper-and-pencil test. They interpreted their results as indicating that "the [aggression] 'suppressing' effect of the SD factor occurs primarily in situations in which the culturally acceptable response is not evident" (p. 532). While Allison and Hunt are careful to refer to social desirability as a "factor," their recasting of Edwards' concept as a "process" perhaps related to "other-directedness" implies a motivational usage similar to that of the present research.

of culturally approved statements to himself and the denial of culturally unacceptable traits. Most importantly, however, to assess the strength of social desirability motivation in a test situation one must be able to determine the actual presence or absence of the traits, characteristics, or symptoms that are denied by the individual. Clearly, a need for approval would not necessarily be implied by the failure to attribute socially disapproved characteristics to oneself when these characteristics are not actually descriptive of the individual. In the development of the M-C scale, a psychometric model was employed which avoids the ambiguities arising from the failure to consider the actual incidence of traits represented in the test items. Items were selected for the M-C SDS from a defined universe representing behaviors which are culturally sanctioned and approved but which are improbable of occurrence.

A low need for social approval implies a degree of independence of cultural definitions of acceptable behavior. The person less motivated by a need for social approval might, in a testing situation, acknowledge certain symptoms, reject them as personally irrelevant, or present other test responses depending on such factors as the strength of his present needs, the kinds of responses required, and the nature of the test stimuli.

The present need construct clearly implies that "social desirability" has considerable generality beyond self-evaluative or test situations, and this study was undertaken to assess the construct's utility for predicting individual differences in response to perceived cultural definitions of appropriate behavior. A situation was required that would be perceived by Ss as demanding of certain socially acceptable behavior. If Ss were presented with a boring, repetitive task and required to perform it for a considerable period of time, it seems probable that frustration would ensue and that negative attitudes would be expressed towards the task. Were this boring task to be presented by an experimenter (*E*) who conspicuously played the role of university professor, authority figure, and omniscient psychologist in the presentation of the experiment and the elicitation of attitudes towards it, Ss with high social approval needs might

be expected to express more favorable (socially appropriate) attitudes than Ss less motivated for approval. The spool packing task used by Festinger and Carlsmith (1959) seemed ideally suited for this purpose and, in slightly modified form, the attitude questionnaire employed by them was deemed equally adequate.

The definition of social desirability as a need for social approval and the belief that this can be attained by means of culturally acceptable behaviors would appear to overlap in some degree with the variable of conformity, and from the present definition of social desirability a relation with conformity would be predicted. The two concepts can be differentiated, however, in that the need for social approval is a *motivational* variable, while conformity refers to a *class of behaviors*. Prediction of a relationship between social desirability and conformity assumes that conformity constitutes a category of behaviors available to individuals seeking to gratify social approval needs. As regards this experiment, there is, nevertheless, a crucial question: would the two concepts differ in their utility for predicting the same behaviors? As a test of this the Independence of Judgment Scale (Barron, 1953), a paper-and-pencil measure of conformity, was included in the experiment to assess its value for predicting attitudes towards the spool packing task.

Finally, since the present construct and its derived test differ from other definitions and measures of social desirability, the same results would not be expected from other social desirability scales. Accordingly, Edwards (1957) SDS was incorporated in the present design to determine its ability to predict the favorability of attitudes towards spool packing.

METHOD

Subjects

Fifty-seven undergraduate male students in introductory psychology classes participated on a voluntary basis in the experiment. The experiment was conducted at the University of Kentucky (26 Ss) and at Ohio State University where 31 Ss were obtained.

Procedure

The experimental procedure was identical at the two universities except for the use of a different *E*

at each institution, and the administration of the Edwards SDS to 29 Ss at Ohio State only. The Ss were individually administered the spool packing task, a four-item questionnaire intended to elicit attitudes towards the packing task, the M-C SDS, the Barron Independence of Judgment Scale and the Edwards SDS. Throughout the entire procedure, *E* maintained a professional and somewhat aloof manner, avoiding any conversation with *S* other than that necessary to conduct the experiment. The following instructions were read to the *S* who was seated at a table directly opposite *E*:

My name is Dr. _____. I'm a psychologist and I'm conducting an experiment on measures of performance. Before we get started on the experiment, I would like you to fill out these questionnaires. Sign your name on all of them.

The Ss then completed the following scales:

1. The M-C SDS, which consists of 33 items with true or false response categories.² An illustrative item is: "I never hesitate to go out of my way to help someone in trouble."

2. Immediately following the M-C scale, *S* completed the Barron Independence of Judgment Scale, a 22-item questionnaire previously shown to be valid for discriminating male conformists from male non-conformists in an "Asch-type" situation (Barron, 1953; Tuddenham, 1958). An illustrative item is: "It is easy for me to take orders and do what I am told."

3. At Ohio State University, 29 Ss completed the Edwards SDS after the Barron scale. The 39 items comprising the Edwards scale were obtained from the MMPI *L*, *F*, and *K* scales and from the Taylor Manifest Anxiety Scale.

When *S* completed the last scale, he was told:

Now for the experiment itself. The materials are this box and the 12 spools. I want you to take these spools, one at a time, and place them in the box. When you are finished, empty the box and refill it one spool at a time. Continue to fill and empty the box until I tell you to stop. Use one hand, and work at your own preferred speed.

S then packed and unpacked the box for 25 minutes while *E* held a stopwatch and pad, conspicuously pretending to be busily engaged in making notes on *S*'s performance. After 25 minutes, *E* said,

O.K., that's all we have in the experiment itself. I hope you enjoyed it. You get a chance to see how you react to the task and so forth. I would like to know what your personal reactions are to the task and the experiment. Would you answer this questionnaire?

The *S* then rated his reactions to the experiment by answering the following four questions, taken from Festinger and Carlsmith (1959):

² A complete description of the M-C scale may be found in Crowne and Marlowe (1960).

1. Was the task interesting and enjoyable? Would you rate how you feel about the task on the scale below where -5 means extremely dull and boring, +5 means the task was extremely interesting and enjoyable, and 0 means the task was neutral, neither interesting nor uninteresting.

2. Did the experiment give you an opportunity to learn about your abilities and skills? Rate how you feel about this on a scale from 0 to 10 where 0 means you learned nothing and 10 means you learned a great deal.

3. From what you know about the experiment, and the task involved in it, would you say the experiment was measuring anything important? That is, do you think the results may have scientific value? Rate your opinion on this matter on a scale from 0 to 10 where 0 means the results have no scientific value or importance and 10 means they have a great deal of value and importance.

4. Would you have any desire to participate in another similar experiment? Rate your desire to participate in a similar experiment again on a scale from -5 to +5, where -5 means you would definitely dislike to participate again, +5 means you would definitely like to participate again, and 0 means you have no particular feeling about it one way or the other.

The three scales, the spool packing task, and the spool packing questionnaire were presented in two orders for the purpose of controlling the possible influence of a sequence effect. Half of the Ss packed the spools first, answered the four questions, and then completed the various scales, while the other half completed the three scales first and were then administered the task and the four questions. The instructions to *S* were modified in accord with the order of presentation used. Ss in the two conditions did not differ significantly with respect to means or variances on any of the measures, and the data were therefore analyzed without regard for the order in which the tasks were presented.

RESULTS

As an initial step, the Ohio State and Kentucky Ss were compared with respect to means and variances on all the measures. No significant differences were obtained and the final analysis of the data was therefore based on the combined *N* of 57. It should be noted that significant results similar to these to be reported below were obtained when statistical analyses were carried out separately for the Kentucky and Ohio State samples. Thus, the findings that follow represent, in essence, the pooled results of a replicated experiment.

In the major hypothesis of the study, it was predicted that individuals with a high need for social approval would express more

TABLE 1
DIFFERENCES BETWEEN HIGH AND LOW M-C
SD GROUPS IN EXPRESSED ATTITUDES

Question	High (<i>N</i> = 30) Mean	Low (<i>N</i> = 27) Mean	Diff.	<i>t</i>
How enjoyable tasks were (rated from -5 to +5)	2.17	-.70	2.87	3.53**
How much they learned (rated from 0 to 10)	5.37	3.22	2.15	2.63**
Scientific importance (rated from 0 to 10)	7.37	5.67	1.70	2.41**
Participate in similar experiment (rated from -5 to +5)	3.63	1.67	1.96	2.57**

** $p < .01$; one-tailed test.

favorable attitudes towards the spool packing task than Ss whose needs for social approval are relatively weaker. To test this hypothesis, Ss' scores on the M-C SDS were dichotomized at the mean (14.93) to yield a high SD group of 30 Ss, and a low SD group of 27 Ss. Scores of the high group ranged from 15-29, while those of the low group were from 5-14. The differences between the mean ratings given to the four attitude questions by the two groups were tested for significance by means of *t*. The results of this analysis are contained in Table 1.

Inspection of Table 1 indicates that the two groups differed significantly in mean ratings on each of the four questions. These differences are all in the predicted direction with the high SD group expressing significantly more favorable attitudes towards the experimental task than the low group. These findings support the general hypothesis that individuals with a strong need for social approval are significantly more likely to express attitudes congruent with perceived situational demands than individuals with a lesser need for social approval.

To assess the relationship between scores on the Barron Independence of Judgment Scale and attitudes towards the spool packing task, an analysis similar to that carried out for M-C SDS was performed. Scores on the Barron scale were dichotomized at the mean (10.39) to yield a high conformity group (*N* = 31), and a low conformity group (*N* = 26). Scores in the low group ranged from 2 to 10, while scores in the high group ranged

from 11 to 20. The mean ratings given to the four questions by the two groups were then compared.

The findings reported in Table 2 indicate only one significant difference between the mean ratings given by the two groups. On Question 2, the ratings given by the high conformity group as to how much they learned about their abilities and skills were significantly higher ($t = 2.02$, $p < .05$) than the ratings assigned by the low conformity group. This single significant difference out of a possible four, indicates that the conformity variable has limited utility for differentiating individuals in the favorability of expressed attitudes towards the spool packing task.

The Edwards SDS, purported to be a measure of test-taking defensiveness—i.e., a measure of a non-test-relevant response determinant—was included in the study as a contrast to the present motivationally defined construct. Scores on the Edwards scale were dichotomized at the mean (32.34), and a high group containing 14 Ss (range of 33-39) and a low group containing 15 Ss (range of 24-31) were obtained.

The significance of the differences between the mean ratings given by the high and low Edwards SD groups to the four questions was also measured by *t* tests. Table 3 presents these data, and indicates that no significant differences were obtained, with the four *t*'s clustering around a value of 0. Quite clearly, social desirability as measured by the Edwards scale is unrelated to attitudes towards the experimental task.

TABLE 2
DIFFERENCES BETWEEN HIGH AND LOW CON-
FORMITY GROUPS IN EXPRESSED ATTITUDES

Question	High (<i>N</i> = 31) Mean	Low (<i>N</i> = 26) Mean	Diff.	<i>t</i>
How enjoyable tasks were (rated from -5 to +5)	1.31	.39	.92	1.04
How much they learned (rated from 0 to 10)	5.27	3.58	1.69	2.02*
Scientific importance (rated from 0 to 10)	6.58	6.55	.03	.04
Participate in similar experiment (rated from -5 to +5)	3.19	2.29	.90	1.13

* $p < .05$; one-tailed test.

As a final step in the analysis of the data, the intercorrelations between the M-C, Edwards, and Barron scales were computed to determine the extent to which scores on these scales are related to each other. Table 4 contains the results of this analysis.

Inspection of Table 4 reveals that M-C SD scores are significantly correlated with both Edwards SD scores ($r = .56$, $N = 29$) and with conformity scores ($r = -.54$, $N = 57$). Scores on the Edwards scale are uncorrelated with scores on the Barron scale ($r = -.12$, $N = 29$). We may conclude that individuals with a high need for social approval (M-C SD) tend to deny the symptoms and complaints represented in the Edwards SD items, and that a high need for social approval is also characteristic of individuals who give responses on the Barron scale indicative of a relative lack of independence of judgment.

DISCUSSION

The major purpose of this study was to assess the utility of treating the construct of social desirability as a motivational variable applicable over a range of situations, in contrast to the usual approach of employing measures of social desirability solely to account for non-test-relevant response variance on personality questionnaires. That social desirability scales designed to measure a specific test-taking attitude can account for a portion of the variance in responses to personality tests has been amply demonstrated (Edwards, 1957; Fordyce, 1956; Wiggins, 1959). There has been a general failure, how-

TABLE 4

CORRELATIONS BETWEEN M-C SD, EDWARDS SD, AND CONFORMITY SCALES

	Edwards SD	M-C SD
Conformity	-.12 ($N = 29$)	-.54** ($N = 57$)
M-C SD	.56** ($N = 29$)	

** $p < .01$.

ever, to consider the possibility that the disposition to dissimulate in a test situation may be an expression of a generalized need to seek social approval. The findings of this study provide clear support for a theoretical rationale which views social desirability in motivational terms, regarding it as a need for social approval accompanied by a belief or expectancy that this need can be satisfied by engaging in culturally and situationally sanctioned behaviors.

As predicted, the attitudes of the high M-C SD groups were significantly and uniformly more favorable toward the experiment than those of the low M-C SD group. We would suggest that the *Es*, as a consequence of their prestige and mildly authoritative manner (reflected in their title, occupation, and behavioral aloofness), were perceived by the high M-C SD *Ss* as persons whose favor was worth courting. Consequently, these high M-C SD individuals were strongly motivated to yield to the demands of the situation: i.e., to tell the *E* that his experiment was interesting, important, personally informative, and worth returning to. In contrast, individuals less strongly motivated for social approval were better able to resist stating what seemed socially appropriate and to offer instead more realistic appraisals of the experiment. Presumably, the less favorable opinions of the low M-C SD *Ss* reflect, in part, the greater freedom of this group from social pressures in the formulation and expression of their opinions. The significant correlation of $-.54$ obtained between M-C SD and conformity would seem to support this formulation. Scores on the Barron scale, however, did not serve to discriminate the favorability of expressed attitudes towards the boring task as well as M-C SD scores. Although one might

TABLE 3
DIFFERENCES BETWEEN HIGH AND LOW
EDWARDS SD GROUPS IN EXPRESSED
ATTITUDES

Question	High ($N = 14$) Mean	Low ($N = 15$) Mean	Diff.	<i>t</i>
How enjoyable tasks were (rated from -5 to +5)	.36	.47	-.11	.08
How much they learned (rated from 0 to 10)	3.79	3.93	-.14	.12
Scientific Importance (rated from 0 to 10)	6.00	5.87	.13	.11
Participate in similar experiment (rated from -5 to +5)	2.07	2.07	0	0

find certain similarities at a definitional level, we would conclude that the need for social approval and conformity (as measured by the Barron scale) are not by any means identical concepts. In terms of the present experimental evidence, conformity is perhaps best conceptualized as defining a class or mode of behaviors in which individuals with a strong need for social approval may engage in a particular situation.

The Edwards scale had no utility whatsoever for predicting differences in attitudes towards the experiment. In a situation where self-evaluation is not a relevant factor, the Edwards scale appears to be of little value in the understanding of motivational determinants of behavior. This is hardly surprising when one recalls that the items included in the Edwards scale refer almost exclusively to the presence or absence of symptoms and complaints, with a consequent restriction of the behaviors that are represented in the item content. By way of contrast, items for the M-C SDS were selected with the intent that they be referents of a construct explicitly defined in motivational terms.

Moreover, scores on the Edwards scale were uncorrelated with conformity scores ($r = -.12$), a finding which suggests, when added to other data recently reported (Wiggins, 1959; Wiggins & Rumrill, 1959), that the Edwards scale may not be a "pure" measure of test-taking attitudes. To date, very high correlations have been reported between the Edwards scale and various MMPI scales and between the Edwards scale and the Taylor Manifest Anxiety Scale (Crowne & Marlowe, 1960; Edwards, 1957; Wiggins, 1959). In contrast to these findings, considerably smaller correlations have been reported between the Edwards scale and tests less related to personal adjustment (Crowne & Marlowe, 1960). It seems reasonable to suggest that the Edwards scale measures the extent to which an individual is willing to admit to symptoms indicative of maladjustment. Thus, we may expect substantial relationships between the Edwards scale and other measures when there is a corresponding overlap in item content (particularly that related to psychopathology).

The present study may be viewed as an attempt to delineate elements in the nomologi-

cal net surrounding a defined construct of social desirability. It seems quite apparent that the "meanings" which may be attached to the Edwards scale as a measure of social desirability are limited in scope and differ in major respects from the demonstrated and implied meanings of the present conception. The findings with respect to the need for social approval strongly support the hypothetical properties ascribed to it. As Cronbach and Meehl (1955) have noted, successful predictions with diverse criteria support the claim of construct validity more forceably than do predictions involving very similar behaviors.

SUMMARY

An attempt was made to assess the utility of defining the construct of social desirability, in motivational terms, as a need for social approval. A new social desirability scale previously developed to measure this variable was administered to subjects at two universities. For comparative purposes, the Edwards Social Desirability Scale and the Barron Independence of Judgment Scale were also included in the study.

Subjects performed a boring task for 25 minutes, and then rated their attitudes towards the experiment. The major hypothesis of the study predicted that individuals with a strong need for social approval would express significantly more favorable attitudes towards the experiment than individuals with a relatively weak need for social approval. The significant findings reported confirmed this prediction. Scores on the Edwards and Barron scales were not significantly related to the favorability of the subject's attitudes.

The overall results were interpreted as contributing to the delineation of the properties which may be attached to two current definitions of the social desirability variable.

REFERENCES

- ALLISON, J., & HUNT, D. E. Social desirability and the expression of aggression under varying conditions of frustration. *J. consult. Psychol.*, 1959, 23, 528-532.
- BARRON, F. Some personality correlates of independence of judgment. *J. Pers.*, 1953, 21, 287-297.
- COWEN, E. L., & TONGAS, P. N. The social desirability of trait descriptive terms: Applications to a self-concept inventory. *J. consult. Psychol.*, 1959, 23, 361-365.

- CRONBACH, L. J., & MEEHL, P. E. Construct validity in psychological tests. *Psychol. Bull.*, 1955, **52**, 281-302.
- CROWNE, D. P., & MARLOWE, D. A new scale of social desirability independent of psychopathology. *J. consult. Psychol.*, 1960, **24**, 349-354.
- EDWARDS, A. L. The relationship between judged desirability of a trait and the probability that the trait will be endorsed. *J. appl. Psychol.*, 1953, **37**, 90-93.
- EDWARDS, A. L. *The social desirability variable in personality assessment and research*. New York: Dryden, 1957.
- FESTINGER, L., & CARLSMITH, J. M. Cognitive consequences of forced compliance. *J. abnorm. soc. Psychol.*, 1959, **58**, 203-210.
- FORDYCE, W. E. Social desirability in the MMPI. *J. consult. Psychol.*, 1956, **20**, 171-175.
- HANLEY, C. Deriving a measure of test-taking defensiveness. *J. consult. Psychol.*, 1957, **21**, 391-397.
- JACKSON, D. N., & MESSICK, S. Content and style in personality assessment. *Psychol. Bull.*, 1958, **55**, 243-252.
- ROSEN, E. Self-appraisal, personal desirability, and perceived social desirability of personality traits. *J. abnorm. soc. Psychol.*, 1956, **52**, 151-158.
- TUDDENHAM, R. D. Studies in conformity and yielding: VII. Some correlates of yielding to a distorted group norm. *ONR tech. Rep.*, 1958, No. 8. (Contract NR 170-159)
- WIGGINS, J. S. Interrelationships among MMPI measures of dissimulation under standard and social desirability instructions. *J. consult. Psychol.*, 1959, **23**, 419-427.
- WIGGINS, J. S., & RUMRILL, C. Social desirability in the MMPI and Welsh's factor scales A and R. *J. consult. Psychol.*, 1959, **23**, 100-106.

(Received February 11, 1960)

ATTITUDES TOWARD SEX ROLES AND FEELINGS OF ADEQUACY IN HOMOSEXUAL MALES¹

BRENDA A. DICKEY

University of Colorado

Many writers have seen homosexuality as either a concomitant or a cause of maladjustment and neurosis, although a few theorists have felt that homosexuals may be either adjusted or maladjusted. Evelyn Hooker's (1957) empirical results suggest that there is some justification for thinking that homosexual males may vary in their degree of adjustment. A major assumption underlying this research is that homosexuals, as well as heterosexuals, may, indeed, be differentially adjusted, and this study attempts to investigate factors that may be associated with varying degrees of adjustment in the homosexual male.

Factors which may be important are suggested by Bennett and by Hooker. Bennett (1947) emphasizes the fact that the heterosexual majority in a disapproving society tends to increase the homosexual's sense of isolation through its selective treatment of him. The heterosexual is able to choose freely his social isolation; the homosexual is not. The only emotional surcease he may find is in the company of others like himself. In another article, Hooker (1956) agrees with Bennett that the homosexual male's adjustment may be greatly facilitated by association with others like himself, and by adopting the standards of the homosexual group. Following this line of thinking, one might be led to expect that the person who has many homosexual contacts or associations would be the most satisfied with his status. The opposite would be expected of the homosexual male who is forced to maintain heterosexual

contacts and associations. This reasoning is subsumed here under a "role conflict" concept; role conflict can be said to exist when the individual is required by the social demands of a particular situation to behave in a manner incongruent with his normal, self-accepting role—assuming, of course, that he does accept the role of homosexuality for himself. Conflicts of this nature might be expected to exist, for example, in an individual who is employed in a job where he must display the characteristics of a typical heterosexual male.

Besides this sort of frustration and thwarting that the homosexual male may have to face in a very real, objective sense, he may also face conflicts which arise in his perceptions of his role. In order to continue building a set of logically consistent hypotheses, this study made the assumption that the person feels most comfortable and assured if he identifies with the role of the typical homosexual male, whatever he may perceive this role to be. A homosexual can therefore face a subjective, cognitive role conflict if he perceives a greater discrepancy between himself and the typical homosexual male than between himself and the typical heterosexual male. This individual is one who feels that the qualities and attributes he possesses are closer to those characteristic of the typical heterosexual male than those characteristic of the typical homosexual male.

The terms "adjustment" and "maladjustment" can have a variety of meanings. Perhaps more satisfactory operations for these terms could be provided by other measures, but these were precluded by the scope of the present study. Instead, subjective measures, relating to the person's feelings about himself—his status, his sense of well-being, ade-

¹ This paper is based upon a master's thesis submitted to the University of Colorado and was partially supported by a grant from the Graduate School of that institution. The data for it were collected while the author was a United States Public Health Fellow under Training Grant M-6613.

quacy, etc.—were used. In order to avoid confusion in terminology, future reference will be made to “feelings of adequacy” instead of to “adjustment.”

In this research, it is expected that feelings of adequacy will be greater in those who:

1. Have homosexual contacts and associations (“Contacts and associations” here include leisure-time activity, “homosexual marriage,” and membership in homosexual groups and organizations.)

2. Suffer fewer pressures toward heterosexual behavior and attitudes (“Pressures toward heterosexual behavior and attitudes” are presumed to be found in masculine or conflictful types of employment. A further indication of pressures is also inferred from an individual’s willingness to reveal his homosexual status to friends, relatives, and work associates, and nonpreferred contact with heterosexuals.)

3. Identify with the typical homosexual male

4. Perceive fewer desirable characteristics in the role of the typical heterosexual male

5. Perceive a smaller discrepancy between themselves and the typical homosexual male than between themselves and the typical heterosexual male

It should be noted here that the conviction with which these hypotheses were made was tempered by the fact that little research in this area has been done, and that Hooker and Hy-Bennett’s suggestions, upon which only Hypothesis 1 is based, do not rest on a large body of experimental evidence. Failure to verify the hypotheses should not, therefore, be necessarily construed as a failure or shortcoming of the admittedly naive ideas and theories on which they are based.

PROCEDURE

Subjects. A total of 47 subjects (Ss) was used in this study. The larger portion was contacted through the cooperation of the Denver and the San Francisco Chapters of the Mattachine Society, a national organization concerned with problems of sexual adjustment.² Anonymity of the Ss was preserved by nameless, numbered questionnaires. In order to facilitate retesting of some Ss at a later date, a contact indi-

vidual in each of these cities maintained a list of Ss and their corresponding questionnaire numbers. Due to the nature of the sampling problems involved in this type of research, the representativeness of the sample is probably as good as can be achieved under the present circumstances. The contacts were instructed to sample as many diverse elements of their respective homosexual populations as possible in order to further the aim of representativeness.

For the purpose of this study, the criterion of homosexuality was defined by the Ss themselves. They were deemed “homosexual” if they were willing to label themselves as such to the contact person. Some individuals volunteered the information that they were predominantly bisexual, or preferred to think of themselves in this manner. However, as far as the investigator was able to ascertain, no S claimed to enjoy sexual relations with the opposite sex *more* than he enjoyed sexual relations with the same sex.

Ss ranged in age from 21 to 63 years, with a mean of 34 years. They had been aware of their own homosexuality from 2 to 45 years, and the mean time they have been aware of their status is 17 years. Eleven considered themselves “homosexually married,” but only one was heterosexually married. A great diversity of occupations was represented, and the residences of the Ss also varied.

The data for this study were collected by means of a paper-and-pencil questionnaire. Fifteen Ss were group tested in Denver at a special meeting of the Mattachine Society, which publicized the session as widely as possible. At this time other blank questionnaires were handed out to interested individuals who thought they could contact friends, acquaintances, etc. who would be willing to complete the test and forward it directly to the writer. Stamped, self-addressed envelopes were provided for this purpose so that the completed questionnaire need not pass through the hands of a third party.

Another 12 questionnaires were individually administered in San Francisco through the aid of the contact who is employed as a minister-counselor in the Mattachine office of that city. He endeavored to give the test to the first 12 homosexuals who came to his office. The remaining 20 questionnaires were returned from various sources, presumably filled out by friends or acquaintances of the original group of 15.

Measurement of Feelings of Adequacy. Feelings of adequacy were measured by the following two devices:

1. The self-ideal discrepancy. S was asked to rate each of 46 traits according to how well it described “himself as he is now,” on a seven-point scale with 1 indicating “exactly like.” This set of ratings described the self-concept, and the same procedure was followed in ascertaining the ideal-self-concept, except that S was asked to rate each of the same words according to how he “would like to be.” It was assumed that the larger the sum of the absolute discrepancies in ratings of these traits under the two different sets, the greater the feelings of inadequacy in the S. Examples of these traits and the rationale

² The author is grateful to the members and friends of this society for their cooperation, and to William A. Scott and Evelyn Hooker for their criticism.

TABLE 1
COEFFICIENTS OF STABILITY
($N = 19$)

Measure	r
Self-Het. Discrepancy	.62
Ideal-Het. Discrepancy	.77
Ideal-Hom. Discrepancy	.75
Self-Hom. Discrepancy	.80
Ideal-Self Discrepancy	.86
Direct Measure of Self-Adequacy	.78

for selecting the particular set used are described in more detail in the section on "Measurement of Subjective Role Conflict."

2. The direct measure. Twenty statements of the MMPI type, referring to S's feelings of adequacy, were prepared and tested for homogeneity on a pilot sample of 45 General Psychology students. S was asked to rate each of the statements on a seven-point scale according to how much of the time he thought it applied to himself. Examples of these statements are: "I am entirely self-confident," "I certainly feel useless," and "I feel that I am a stable person."

Measurement of Objective Role Conflict. Objective role conflict was said to exist for a homosexual male employed in a job where he must display characteristics of a typical heterosexual male. Thus this study assessed S's occupations, and these were rated by the investigator and another judge according to whether or not they seemed likely to pose conflict for the individual. "Conflict" was defined by two criteria: requiring the individual to behave in a manner characteristic of a typical heterosexual male, and assumption of this role in a type of employment which also required frequent contact with a predominantly heterosexual public. A coding of 1 was made to indicate absence of conflict, 2 indicated that an ambiguous or neutral occupation was held, and 3 indicated the likelihood of sex role conflict. A self-employed bookshop operator, a student, and an artist were, for example, rated 1; and a railroad engineer, an engineering geologist, and a lawyer were rated 3.

Measurement of Subjective Role Conflict. These measures were developed from lists of the same traits as were used to determine the self-ideal discrepancy. In other parts of the questionnaire S was asked to rate each of the 46 traits according to how well it described the "typical male homosexual," the "typical female heterosexual," and the "typical male heterosexual." The same seven-point scale was utilized as before. The rationale for selecting the specific traits used rests upon articles like that of Parsons (1956), wherein some suggestions are made concerning differentiating qualities assigned to men and women in this culture, and upon intuitive hunches regarding traits which may be thought of as characteristic of men, women, and male homosexuals. Some of the

descriptive adjectives were fillers, but the majority were selected to represent diverse areas of human behavior, regarded in popular stereotypes as characteristic of these three groups. Some illustrative examples are: "able to get along with everybody," "aggressive," "ambitious," "creative," "flighty," "fault finding," "intellectual," "irresponsible," "mature," "self-centered," "sociable," and "talented."

Reliability of the Measures. Nineteen of the original Ss were recontacted and administered a shortened form of the questionnaire approximately 3.5 to 4 months after the original test. Coefficients of stability were then calculated for the discrepancies between the self-concept and the typical heterosexual male, the ideal-self and the typical heterosexual male, the ideal-self and the typical homosexual male, the self and the typical homosexual male, and the ideal-self and the self. A coefficient of stability was also calculated for the direct measure of adequacy. These product-moment correlations are shown in Table 1.

Split-half reliability coefficients were calculated on the total original sample for the two measures of adequacy. These, corrected for length by the Spearman-Brown formula, are .96 for the self-ideal discrepancy, and .91 for the direct measure.

Indications of the Validity of the Measures. An indication of the validity of the two measures of self-adequacy was arrived at by correlating the scores between the direct measure and the self-ideal discrepancy. The obtained product-moment r of .53 was judged sufficient to allow the use of both as representing feelings of adequacy in this study.

Some support for the validity of certain of the other discrepancy measures is also provided by data

TABLE 2
MEAN RAW SCORE DISCREPANCIES FOR THE SELF LESS THE TYPICAL HOMOSEXUAL MALE, HETEROSEXUAL FEMALE, AND HETEROSEXUAL MALE. MEAN RAW SCORE DISCREPANCIES FOR THE TYPICAL HOMOSEXUAL MALE LESS THE TYPICAL HETEROSEXUAL FEMALE AND HETEROSEXUAL MALE

Discrepancy	Mean	SD	t^a
Self-Hom.	61.6	21.7	
Self-Fem.	68.3	23.9	-2.79*
Self-Het.	75.8	27.8	-3.96**
Hom.-Fem.	57.6	25.6	
Hom.-Het.	73.2	28.2	-6.26**

Note. - Total $N = 47$. One S was excluded from the Hom-Het analysis because he obtained a discrepancy score of 0. It is assumed that he misunderstood the instructions, so that the ratings he gave to the typical homosexual male were duplicated on the page which asked for ratings on the typical heterosexual male.

^a t of correlated differences.

* $p < .01$, two-tailed test.

** $p < .001$, two-tailed test.

from the original sample of homosexuals. Mean raw score discrepancies were calculated for the following: the self-typical homosexual male (self-hom.), the self-typical heterosexual female (self-fem.), the self-typical heterosexual male (self-het.), the typical homosexual male-typical heterosexual female (hom-fem.), and the typical homosexual male-typical heterosexual male (hom-het.). These differences between mean discrepancies were then tested for significance (Table 2). It is readily apparent that the average self is seen as more like the typical homosexual male than like the typical heterosexual female, and more like the typical heterosexual female than like the typical heterosexual male. It is also apparent that the typical homosexual male is perceived as more like the typical heterosexual female than like the typical heterosexual male. These results correspond with common assumptions concerning the relative similarities of homosexuals and typical males and typical females; hence, they lend a certain degree of confidence to the discrepancy measures on which they are based.

RESULTS

The major portion of the data analysis was carried out in the following manner. The raw self-adequacy scores on both measures and the various raw discrepancy scores were ordered in a frequency distribution and inspected to find equal intervals which would cover all the distributions of the total number of discrepancies to be used in the analysis. It was decided to use six intervals, in order to facilitate punching the data into a single IBM card column, and yet allow a sizeable number of Ss in each group. Unless otherwise noted, all analyses to be reported in this section are based on these group scores rather than on the raw scores. It should also be noted (a) that the smaller the group scores, e.g., 1 as opposed to 6, the smaller the absolute discrepancy, and (b) that in the case of the self-adequacy measures, the smaller the group scores, the greater the feelings of adequacy.

Results Relating to Objective Role Conflict

Hypothesis 1 stated that the more homosexual contacts and associations participated in by the S, the greater would be his feelings of adequacy. To test this, the following operations were performed. Mean feelings of adequacy on both measures were calculated for:

1. Ss associating predominantly with other

TABLE 3

MEAN FEELINGS OF ADEQUACY FOR SUBJECTS WHO ARE EITHER "HOMOSEXUALLY MARRIED" OR "HOMOSEXUALLY UNMARRIED"

	<i>N</i>	Mean	<i>SD</i>	<i>t</i>
Direct Measure				
Homosexually Married	11	2.82	1.47	1.149
Homosexually Unmarried	36	3.36	1.02	
Self-Ideal Discrepancy				
Homosexually Married	11	2.18	1.33	2.956*
Homosexually Unmarried	35	3.51	1.27	

* $p < .01$, two-tailed test.

homosexuals and those associating predominantly with heterosexuals

2. Ss belonging to homosexual social groups or organizations, and those belonging to neither

3. Ss who considered themselves homosexually married and those who did not

No significant differences were found in the first two operations; however, the differences between the mean self-adequacy scores reached significance on one measure in the third operation (Table 3). Apparently homosexual males can be said to feel more adequate if they are homosexually married.

To test Hypothesis 2 that feelings of adequacy will be greater in Ss who suffer fewer pressures toward heterosexual behavior and attitudes, the following operations were performed. Mean feelings of adequacy on both measures were calculated for:

1. Ss who were rated as being in nonconflictful jobs, ambiguous jobs, and conflictful jobs

2. Ss expressing much satisfaction (rating of 1) with nonconflictful jobs, ambiguous jobs, conflict jobs, and for Ss expressing less satisfaction (ratings of 2, 3, and 4) with the three job categories

3. Ss expressing preference for contact with homosexuals and actual leisure-time association predominantly with heterosexuals, preference for contact with homosexuals and actual leisure-time association predominantly with homosexuals, preference for contact with heterosexuals and actual leisure-time association predominantly with heterosexuals, and

then takes on two meanings: "deviancy" in the sense that all Ss have departed from the cultural norm of preferring a heterosexual female for a sex partner; and "deviancy" in the sense that some Ss have also chosen to reject the prescribed sex role—that of the typical heterosexual male.

Since a comparison of mean self-adequacy scores on the direct measure for this homosexual male sample vs. a random sample of male students at the University of Colorado³ had yielded no statistically significant difference, we are led to conclude that the former meaning does not necessarily lead to feelings of inadequacy, while the latter meaning does, since this research has finally found that those homosexual males who see the prescribed sex role as uncongenial are those who are also inadequate. But those homosexual males who do adhere to the cultural standards of feeling, perceiving, emulating, and idealizing the typical heterosexual male are more likely to feel self-satisfied and adequate.

Of course, all of these tentative interpretations must be dealt with cautiously. The sample used in this study was in no way randomly selected, and the generalizability of the results to the total homosexual male population is, therefore, questionable.

SUMMARY

This research was a study of feelings of adequacy in homosexual males. Designed to test a set of simple hypotheses, the study focused on objective and subjective role conflict as possible factors relevant to homosexual males' feeling of adequacy.

Forty-seven homosexual males were anonymously assessed by means of paper-and-pencil questionnaires. Feelings of adequacy were measured by: the discrepancy between the self- and the ideal-self-concepts on a list of 46 traits deemed relevant to the homosexual male's status; and responses to a list of 20

MMPI-type statements presumably reflecting feelings of adequacy. The correlation between these two measures of adequacy was .53. Information relative to objective role conflict was obtained through answers to questions regarding the subject's status from which one could infer pressures toward heterosexual behavior and attitudes. Subjective role conflict was assessed by means of ratings assigned to the same 46 traits as were utilized in the self-ideal discrepancy, with separate ratings obtained for the "typical homosexual male," the "typical heterosexual female," and the "typical heterosexual male."

The findings did not, in general, support the role conflict hypotheses. Instead, the results could more readily be interpreted as indicating that subjectively adequate homosexual males were those who tended to identify with the masculine norms of the dominant culture. Feelings of adequacy were associated with: job satisfaction, preference for leisure-time association with heterosexuals, idealization of the role of the typical heterosexual male, and identification with the typical heterosexual male rather than with the typical homosexual male.

Since this was not a random sample of homosexual males, it is recommended that caution be used in generalizing from these results.

REFERENCES

- BENNETT, E. H. The social aspects of homosexuality. *Med. Pract.*, 1947, 217, 207-210.
- BROWN, D. G. The development of sex-role inversion and homosexuality. *J. Pediat.*, 1957, 50, 613-619.
- HOOKE, EVELYN. A preliminary analysis of group behavior of homosexuals. *J. Psychol.*, 1956, 42, 217-225.
- HOOKE, EVELYN. The adjustment of the male overt homosexual. *J. proj. Tech.*, 1957, 21, 18-31.
- NEWCOMB, T. M. *Social psychology*. New York: Dryden, 1950.
- PARSONS, T. Age and sex in the social structure of the United States. In C. Kluckhohn & H. A. Murray (Eds.), *Personality in nature, society, and culture*. New York: Knopf, 1956. Pp. 363-375.

(Received February 23, 1960)

³ The University of Colorado cross section was studied by Robert Kassebaum and Leon Rappaport, under the direction of William A. Scott.

A SECOND VALIDATION OF A LONG-TERM RORSCHACH PROGNOSTIC INDEX FOR SCHIZOPHRENIC PATIENTS¹

ZYGMUNT A. PIOTROWSKI AND BARRY BRICKLIN

Jefferson Medical College of Philadelphia

In 1952 a group of Rorschach prognostic signs were presented with which the follow-up conditions of schizophrenic patients could be predicted (Piotrowski & Lewis, 1952). An essentially postdictive methodology was used, i.e., the prognostic signs were stated after the follow-up groupings had been formed. In 1958 these signs were applied to 30 schizophrenic patients, some of whom improved over an interval of at least 3 years and some of whom remained unimproved over the same interval. On the basis of this application the signs were revised (in order to increase reliability) resulting in the 1958 prognostic index. The 1958 prognostic index was then validated on a group of 70 schizophrenics (Piotrowski & Bricklin, 1958). The purpose of the current investigation was to test the validity of the index on a group of patients who differed in many important respects from the 70 used in 1958. The prognostic index was again revised slightly, and was revalidated on 103 additional schizophrenic patients (the 1959 group). This slightly revised version was then reapplied to the 1958 sample so as to facilitate comparisons between the two groups. It should be kept in mind that the 1959 revision did not change the cutoff point, and the relative distribution of the 1958 cases has not changed in any way, i.e., the two samples constitute independent validating evidence for the prognostic index.

¹ The authors wish to thank Clellen Morgan and Robert Ballard of the Veterans Administration Regional Office, Philadelphia, and David Cohen of the Veterans Administration Hospital, Coatesville, Pennsylvania, for their invaluable assistance in making this investigation feasible. This study was supported by a grant from the Research Committee of the Supreme Council, thirty-third degree, Scottish Rite Freemasonry, Northern Masonic Jurisdiction.

METHODOLOGY

The essential validating methodology in both the 1958 and the 1959 samples was to compare predictions made on the basis of the Rorschach prognostic index against follow-up statements made on the basis of independent clinical judgments.

In the selection of cases, a rule of inclusion was that there be copious follow-up data on each schizophrenic patient, consisting of psychiatric interviews, psychiatric social worker interviews, interviews of the patient's family members by psychiatric social workers, and staff conference reports. This information had to extend at least 3 years past the time at which a Rorschach test had been administered. It was to this Rorschach test that the prognostic index was applied. The actual year in which the Rorschach had been given was unimportant so long as there was follow-up data at least 3 years subsequent to this year. The other standard of inclusion was that each patient be independently diagnosed as schizophrenic.

The first step in the procedure was to search through the psychological test files and locate all schizophrenic cases to whom Rorschach tests had been administered at least 3 years ago. The second step was to see if there was extensive follow-up data on all such cases which extended at least 3 years subsequent to the time of Rorschach testing. The third step was to insure that each case had maintained the diagnosis of schizophrenia over the entire interval for which information was available. The same clinicians who made the follow-up designations (see below) made all decisions to include or exclude cases. All such decisions were made without knowledge of Rorschach results. Only one-fourth of the Rorschach records taken at least 3 years earlier could be used. The remaining three-fourths of the cases did not have adequate follow-up information.

The methodology consisted of comparing predictions made on the basis of the Rorschach prognostic index against follow-up statements as designated by experienced clinicians working independently, using the clinical and life history data enumerated above (Lewis & Piotrowski). On the basis of this follow-up information each patient was designated as improved or unimproved over the x -year interval, where x was at least 3. The prognostic index was applied to the Rorschach tests each one of which

had been administered at least 3 years prior to the time to which the follow-up extended. All Rorschach tests were identified by number only; the rater (Bricklin) had no other information at his disposal. This design eliminated the possibility of contaminating factors. The rater who applied the prognostic index had no data as to the follow-up conditions of the patients; the clinicians who made the follow-up designations had no Rorschach data. A prediction (improved, unimproved) based on the Rorschach prognostic index was made for each case. A score of +2 or more would predict the patient to remain unimproved; +1 or less would indicate the patient to be improved. These predictions were then confronted with the independent follow-up designations (improved, unimproved) and the two results were compared by the chi square technique.

The follow-up length (at least 3 years) was used not only in order to make possible a meaningful statement as to follow-up condition but to insure the correctness of the initial diagnosis. Since our methodology demanded that there be extensive follow-up data on each patient, the chance of our having included a nonschizophrenic patient was, for all intents and purposes, ruled out.

The 1958 and the 1959 samples of schizophrenic patients were chosen so as to differ from each other in age, intelligence, sex composition, socioeconomic status, and duration of time elapsed between onset of manifest psychosis and Rorschach examination time. The purpose was to validate the prognostic index on as wide a range of schizophrenic patients as possible.

Follow-Up Designations

The same clinical follow-up criteria of improvement and unimprovement were used in both the 1958 and the 1959 groups. Every available source of information—psychiatric interviews and evaluations plus psychiatric social work reports, family reports on patients' adjustment and behavior, and staff conference notes—was scrutinized. The patient had to improve in all three of the following areas to be designated as improved. If he failed to improve in all three areas or grew worse, he was classified as unimproved. The three areas are:

1. Thought processes. The relevance, coherence, sense of reality, comprehensiveness, consistency, confidence, and valid self-criticism in making statements, were considered.

2. Psychosocial relations and work. The capacity of the patient to form meaningful emotional relations with others was considered. The employment record of each patient was analyzed from two standpoints: as an additional check on his capacity to live with others, and to yield information on his ability to do some kind of work reasonably effectively. On hospitalized patients the hospital work history was considered. The patient had to display an increased capacity for productive work and for meaningful and more constructive interhuman relations to be designated as improved in this area.

TABLE 1
POPULATION CHARACTERISTICS OF 1958
GROUP AND 1959 GROUP

Variables Being Compared	1958 Group (N = 70)		1959 Group (N = 103)		<i>t</i>
	Mean	SD	Mean	SD	
Age	28	8.4	34	7.2	4.76*
IQ	118	17.0	97	15.1	8.24*
Follow-up Interval	6.0	3.7	6.5	2.8	1.06

* $p < .01$.

3. Attitude towards self. This refers to the degree of anxiety and to the degree of self-acceptance. To be designated as improved in this area, the patient had to be more comfortable with himself, and had to give evidence of feeling in a realistic way that his life had become less troublesome and less difficult.

It is important that two points be kept in mind when follow-up designations are to be made. (a) Some symptoms of schizophrenia generally persist even in improved cases. There is generally some degree of affective blunting. Some traces of delusional thinking generally persist, even in improved cases. (b) When dealing with one of those patients termed "episodic" by Bleuler (1950), it is important not to consider daily (or even hourly) fluctuation in condition as essential change. This is especially true of manic or depressive mood phases. These mood phases are almost always transient. An inspection of the entire duration of the follow-up interval must be made until its trend is revealed. With episodic patients the frequency of lucid intervals, and the degree of defect shown in the lucid intervals, become the deciding criteria.

1958 Sample

It had been possible to isolate from our psychological test files 70 cases which met the conditions of inclusion, i.e., patient diagnosed schizophrenic, Rorschach test available, follow-up information extending at least 3 years subsequent to Rorschach, diagnosis of schizophrenia maintained over entire interval for which information was available. The condition most difficult to meet in this group as well as in the 1959 group was that of obtaining huge masses of follow-up information so that an accurate appraisal as to the trend of each patient's illness could be made.

As shown in Table 1, the average intellectual level was above average in the 1958 group, with a mean Wechsler-Bellevue IQ of 118, the standard deviation being 17. Nearly all of the patients were first admissions and the Rorschach tests to which the signs were applied had been administered 2 months to 2 years after the onset of manifest psychoses. The dura-

tion of this particular interval was gathered from the follow-up sources listed above and did not necessarily correspond to the duration of hospitalization. The mean age of the 1958 group was 28 years, the standard deviation being 8.4 years. All of the patients in this group belonged to the middle and high middle socioeconomic classes. There were 29 men and 41 women in this sample. The mean follow-up interval (time elapsed between that point at which the Rorschach was administered and the time to which the patient was followed) was 6.0 years, the standard deviation 3.7 years.

1959 Sample

All patients met the conditions of inclusion as outlined previously. These patients were chosen so as to differ in the above mentioned respects from the 1958 group. To satisfy these requirements the patients were selected from a Veterans Administration Regional Office (53 patients) in Philadelphia and from the Veterans Administration Hospital (50 patients) at Coatesville, Pennsylvania. The 103 veterans in this group had become manifestly psychotic from 2 to 10 years prior to the time at which they were administered the Rorschach test. At the time of examination their mean Wechsler-Bellevue IQ was 97, the standard deviation 15.1 as is shown in Table 1. The mean age was 34 years, the standard deviation 7.2. The 1959 group contained 97 men and 6 women. The differences between the 1959 group and the 1958 group in IQ and age are statistically significant ($p < .01$) as are the differences in interval between onset of manifest psychosis and psychological examination, and sex composition. Practically all members of the 1959 sample belonged to a socioeconomic class below the middle class. The mean follow-up interval in this group was 6.5 years, the standard deviation 2.8. The durations of times over which the patients were followed subsequent to their Rorschach tests did not differ significantly in the two groups.

RESULTS

Validation

1958 Sample. There was one disagreement among the two clinicians as to follow-up designations. This difference was reconciled in open conference until one decision was reached for the case (for design purposes).

It had been hypothesized that high index scores would be associated with failure to improve, and low index scores with improvement. On the basis of the previous investigations, +2 and more was chosen as the cutoff point that would distinguish the unimproved patients from the improved patients. In the 1958 validation group (Piotrowski & Bricklin, 1958), 49 patients obtained scores of at least +2 points; 45 of them were worse or did not

TABLE 2
CHI SQUARE ANALYSIS OF PROGNOSTIC INDEX
SCORES OBTAINED IN 1958 GROUP

Index Scores	Follow-up Status		
	Improved	Unimproved	Total
2 and more	4	45	49
1 and less	18	3	21
Total	22	48	70

Note.— $df = 1$, $\chi^2 = 41.03$, $p < .01$.

change during the follow-up period, i.e., 45 of these cases had been independently designated as unimproved. Of the 21 patients with scores of less than +2 points, 18 were independently designated as improved. The prognostic index correctly predicted 90% of the 70 patients' follow-up conditions. As may be noted on Table 2, the chi square value of 41.03 ($df = 1$) indicates that such results could occur by chance less than 1 in 100 times.

1959 Sample. There were two disagreements between the two clinicians as to follow-up designations. As above, these differences were reconciled in open conference until one decision was reached for each case. Seventy-two schizophrenic patients attained scores on the prognostic index of at least +2 points; 70 of these patients had been independently designated as unimproved. Thirty-one patients attained index scores of less than +2 points; 22 of these patients had been designated as improved. The chi square value of 56.66 ($df = 1$) indicates that such an association between the index scores and follow-up conditions could have occurred by chance less than 1 in 100 times. Thus the follow-up conditions of 89% of the patients had been correctly predicted by the prognostic index in the 1959 group (see Table 3).

Comparison of Results

Minor changes have been made in the presentation of the 1959 data (from the 1958 data) in the grouping of several signs and in the weightings of two. Signs 3 and 6, and Signs 5 and 7, were combined because of the relatively low frequency of occurrence of

TABLE 3

CHI SQUARE ANALYSIS OF PROGNOSTIC INDEX
SCORES OBTAINED IN 1959 GROUP

Index Scores	Follow-up Status		Total
	Improved	Unimproved	
2 and more	2	70	72
1 and less	22	9	31
Total	24	79	103

Note.— $df = 1$, $\chi^2 = 56.66$, $p < .01$.

Signs 6 and 7. The weightings of Signs 2 and combined 3 & 6 were lowered from 3 points to 2 points. This revised index was applied to both the 1958 and the 1959 validation groups in order to make the results directly comparable. Neither the original validity of the single signs nor of the original prognostic conclusions in the 1958 group was affected in any way. The same cutoff point of +2 was used with both groups. The weighting changes did not affect the relative distribution of cases in the 1958 sample.

As can be seen, the prognostic index successfully predicted the follow-up conditions of 90% of the patients in the 1958 group, and 89% of those in the 1959 veteran group.

The variables in which the two validation groups differed—intelligence, age, etc.—did not affect the validity of the prognostic index. There is a tendency among the veterans (the 1959 group) to show a somewhat lower incidence of improvement in the low score group. In the 1959 veteran group, of the 31 persons obtaining a prognostic index score of 1 or less, 71% actually were improved. In the 1958 group, of the 21 patients obtaining the same score, 86% improved. This finding, which may be related to differential treatment procedures, remains to be investigated.

The 1959 veteran group was composed of ambulatory or milder VARO cases ($N = 53$), and more severe VAH cases ($N = 50$). The 1958 group ($N = 70$) falls between these two other groups in terms of severity of illness; these patients were hospitalized but were early and mild cases at the time. It is interesting to note that the mean prognostic index scores were 2.8 ($SD = 3.3$) in the VARO

cases; 3.9 ($SD = 4.3$) in the 1958 group; and 5.8 ($SD = 4.0$) in the VAH group. The mean prognostic index scores reflect the increasing severity of defect in the three groups. The difference in the prognostic index scores of VARO and VAH cases is significant at $p < .01$ level; that between the 1958 group and the VAH cases at $p < .02$. The difference in scores between the 1958 group and the VARO cases falls at the $p < .10$ level (t tests).

Two of the signs, 3 & 6 combined and 4, were more frequent in the 1958 group. Thus it is possible that these signs are related to intelligence. The appearance of these signs apparently requires on the part of the patient a critical attitude toward thinking and some facility in verbalizing thoughts, with the concomitant condition that these thoughts and their evaluations be defective. However, the habit itself of thinking and speaking in terms of probabilities rather than certainties ("could be, I don't know"; "might be anything that has a shape"; etc.) is correlated with intelligence and therefore shows up to a greater extent in the brighter 1958 group. Signs 11 (determinant scarcity) and 12 (content mo-

TABLE 4

THE PROGNOSTIC SIGNS, THEIR WEIGHTINGS,
AND CHI SQUARE VALIDITY($N = 173$)

Sign Number	Name of Sign ^a	p
1	Human Movement Responses 0 or 1 and Sum Color Responses outweighs Sum M by at least 3 (4)	<.01
2	Response repetition (perseveration) (2)	<.05
3 & 6	Vagueness of perception and meaning or inappropriate conceptual connection (2)	<.02
4	Indeterminate form responses (2)	<.02
5 & 7	Breakdown of interpretive attitude or blurring of difference between imagination and sensation (2)	<.01
8	Absurdly inconclusive explanations (2)	<.02
9	Absence of human content (2)	<.01
10	$F + \%$ below 60 (2)	<.01
11	Determinant scarcity (1)	<.01
12	Content monotony (1)	<.01
13	No Human Movement Responses (2)	<.01
14	At least 5 Human Movement Responses (-2)	<.01

^a The weightings are given in parentheses. For a more detailed description of each sign, including examples of each, the reader is referred to Piotrowski and Bricklin (1958).

notony) were more frequent in the less bright 1959 veteran group. This would be expected. As a rule, the variety of Rorschach components decreases with decreasing intelligence. By retaining signs which appear with differing frequency in varying populations we are able to apply the same set of signs to different types of schizophrenics.

The prognostic validity of each sign was measured by the chi square technique. This was done for the 1958 and the 1959 groups separately as well as for the entire sample. Using the entire sample of 173 cases, a four-cell contingency table was formed: the improved and unimproved patients formed one dimension, those manifesting and those not manifesting the sign the other dimension. The results are given in Table 4. The following signs differentiated between the improved and unimproved cases at the $p < .01$ level: 1, 5 & 7, 9, 10, 11, 12, 13, and 14. Three signs differentiated at the $p < .02$ level: 3 & 6, 4, and 8.² Sign 2 differentiated at the $p < .05$ level.

Reliability

The reliability of the prognostic index was tested independently of the main study by having five raters apply the index to each of 10 schizophrenic cases, and by having another rater and one of us (Bricklin) independently apply the index to 25 schizophrenic cases.

Five raters were given seven unimproved and three improved schizophrenic Rorschach records. These records were chosen at random from among improved and unimproved cases in proportion to the rate at which improved and unimproved schizophrenic cases appear in the general schizophrenic population. The raters, of course, had no knowledge of this decision rule.

There was no disagreement in the case of eight patients whom all five raters placed in the same group, improved or unimproved. One rater disagreed with the rest of them by placing one unimproved patient among the improved; and one rater disagreed with his fellows by placing an improved patient among

the unimproved. Out of a total of 50 predictions, only 2 were incorrect as to prognostic conclusion. On four patients, all five raters had identical prognostic index scores. On four other patients the greatest difference in scores among all five raters was only two points. On the remaining two patients, the highest and lowest scores differed by four points; however, this difference was critical in only one case by leading to a prognostic conclusion opposite to that of the majority of raters.

In addition, one other rater (Carter Zeleznik) applied the index to 18 unimproved and 7 improved cases chosen in the same manner as above. These same cases were independently scored by one of us. The prognostic index scores differed by two points in four cases, and by one point in one case. In only one case, however, did the difference (of two points) influence the prognostic conclusion. The reliability can be considered satisfactory.

DISCUSSION

Long experience has shown, as Bleuler (1950) noted, that a schizophrenic frequently undergoes marked and unpredictable changes in personality during the first years after the onset of manifest psychosis. Such marked and unpredictable changes are rare 3 or more years after the onset of manifest psychosis. This factor makes long-term prognostic investigations more realistic, at the present time, than short-term investigations. The rapid alternations of condition often so characteristic of the early years of schizophrenia render the problem of making accurate and meaningful follow-up statements most complex. There is a strong tendency for the eventual course of the illness to make itself known after the first 3 years. The frequency of essential changes in condition is exceedingly low after the first 3 years. Another factor which complicates short-term prognostic study is the differential effects of various treatment procedures. Psychotherapy and other therapeutic procedures are much more effective in the beginning of manifest psychosis than in later years when schizophrenics become much less responsive to environmental influences, including therapy (Gottlieb & Huston, 1943). This decrease in personality variability with time favors long-term prog-

² Signs 3, 5, 6, and 7 were all individually valid at at least the $p < .05$ level, as they are in combination.

nostic research and greatly complicates short-term prognostic studies. It may be mentioned that, in this study, differential treatment procedures did not seem to affect the validity of a long-term prognostic index prediction in any consistent manner.

Other attempts to prognosticate the outcome of schizophrenia have taken many courses. Kantor (1953), among others, has differentiated so-called "process schizophrenics" from "reactive schizophrenics," the prognosis of the latter being more favorable than that of the former. The reactive cases are characterized by a normal prepsychotic personality and an acute onset usually accompanied by a "logical" precipitating factor. The reactive cases are also characterized by a clouded sensorium. The process types are best characterized as "thinking disorders" and generally have insidious onsets of disease. In a system generally similar to this, Langfeldt (1937) has distinguished so-called "typical" (process) and "atypical" (reactive) cases.

One may also find in the literature many studies which list clinical symptoms or syndromes of schizophrenia along with the percentage of improved cases associated with each.

Difficulty in using the above mentioned procedures as prognostic tools has to do not with their essential validities, but with the difficulty of adapting them to the individual case. Many of these approaches depend on accurate and reliable case histories which are often impossible to obtain. It may also be noted that clinical signs and syndromes more often than not are highly variable in the individual case.

SUMMARY AND CONCLUSIONS

In 1958 a Rorschach prognostic index for the prediction of a schizophrenic's clinical condition (improved or unimproved) 3 or more years after testing was offered. The index had been validated on a group of 70 followed-up patients (Piotrowski & Bricklin,

1958). The present report describes the second validation of the index on a group of 103 schizophrenic patients, differing from the first patient group in many ways, including average intelligence, age, severity of illness, and distributions of sexes. The results were virtually the same. In 90% of the cases in the first group, and in 89% of the cases in the second group, the prognostic index successfully predicted the outcome conditions of the schizophrenic patients as either improved or unimproved.

Since the implications of a long-range prognostic index which validly differentiates between schizophrenics who will be improved or unimproved are obviously serious, it is advisable to submit the prognostic index to additional tests. It must be remembered that the index applies to schizophrenics, and not to cerebral organic cases or psychoneurotics. The validity of the diagnosis of schizophrenia is an essential factor determining the degree of validity of the index.

REFERENCES

- BLEULER, E. *Dementia praecox or the group of schizophrenics*. New York: International Univer. Press, 1950.
- GOTTLIEB, J. S., & HUSTON, P. E. Treatment of schizophrenia: Follow-up therapy in cases of insulin shock therapy and in control cases. *Arch. Neurol. Psychiat.*, 1943, 49, 266-271.
- KANTOR, R. E., WALLNER, E., & WENDNER, C. L. Process and reactive schizophrenia. *J. consult. Psychol.*, 1953, 17, 157-162.
- LANGFELDT, G. The prognosis in schizophrenia and the factors influencing the course of the disease. *Acta psychiat.*, 1937, Suppl. 13. (See also Suppl. 80, 110, 113)
- PIOTROWSKI, Z. A., & BRICKLIN, B. A long-term prognostic criterion for schizophrenics based on Rorschach data. *Psychiat. Quart. Suppl.*, 1958, 32, 315-329.
- PIOTROWSKI, Z. A., & LEWIS, N. D. C. An experimental criterion for the prognostication of the status of schizophrenics after a three-year interval based on Rorschach data. In P. Hoch & J. Zubin (Eds.), *Relation of psychological tests to psychiatry*. New York: Grune & Stratton, 1952. Pp. 51-72.

(Received February 25, 1960)

SOME EFFECTS OF STIMULUS VARIATION ON SPIRAL AFTEREFFECT IN ORGANIC AND NONORGANIC SUBJECTS¹

RONALD M. SINDBERG²

Duke University

Following the reports by Price and Deabler (1955) and Garrett, Price, and Deabler (1957) of studies in which perception of the negative spiral aftereffect (SAE) was found to discriminate with great accuracy between brain damaged patients and two nonorganic control groups, the clinical implications of this perceptual phenomenon have been studied by many different investigators. Some, such as Gilberstadt, Schein, and Rosen (1958) and Philbrick (1959), have carefully followed the methods of Price and Deabler (1955). Others, such as Gallese (1956), Davids, Goldenberg, and Laufer (1957), Spivak and Levine (1957), and Page, Rakita, Kaplan, and Smith (1957), have used somewhat different procedures and/or scoring systems. While the variations employed in these later studies make it difficult to compare their results precisely, at least one generalization appears to be warranted. While brain damaged subjects (Ss) as a group do tend to report the negative SAE less frequently than do either normal Ss or nonorganic psychiatric patients, the differences between groups are far less clear-cut than originally reported by Price and Deabler.

In trying to explain this discrepancy, several investigators have suggested that the composition of the subject samples employed

in different studies has been a crucial factor. In many cases the brain damaged Ss were older, more chronic patients than were the control groups studied. Location and type of brain damage are very probably other factors of importance. Both Gallese (1956) and Page et al. (1957) noted that many prefrontal lobotomy cases reported the SAE as readily as did normal Ss, and Aaronson (1958) has suggested that involvement of the temporal lobes is especially likely to eliminate the normal SAE response. Other investigators have raised the possibility that the apparent decrements shown by the brain injured may result chiefly from an inability to report the aftereffect rather than from failure to perceive it. Gallese (1956) attempted to deal with this problem by using more directive and probing instructions than did Price and Deabler (1955), while at the same time he revised his scoring procedures to reduce the likelihood that reticence or difficulty in report would contribute to low scores.

In the studies cited above, there have been only minor variations in the stimulus conditions under which the spiral was presented. Although the conditions utilized by Price and Deabler (1955) seem to be close to optimal for perception of the SAE by normal Ss, no attempt has been made to establish the optimal conditions for its perception by the brain injured or for differentiating between brain damaged and nonorganic Ss. The present study was designed to investigate the effects of varying concomitantly two easily controlled and readily quantified stimulus conditions, which, from studies of other aftereffect and apparent motion phenomenon (Hammer, 1949; Teuber & Bender, 1949), seemed

¹ Based on a dissertation submitted in partial fulfillment for the doctor of philosophy degree in the Department of Psychology, Duke University, Durham, North Carolina. The author wishes to thank Gregory Kimble and the other staff members at Duke University, Butner (North Carolina) State Hospital, and Veterans Administration Hospitals at Richmond, Virginia, Fayetteville, North Carolina, and Durham, North Carolina, who cooperated in this study.

² Now at Laboratory of Neurophysiology, Department of Physiology, University of Wisconsin Medical School, Madison, Wisconsin.

TABLE 1
DESCRIPTION OF PATIENTS STUDIED

Cortical Damage Group		NP Control Group	
Diagnosis	Number	Diagnosis	Number
Chronic brain syndrome, associated with trauma	20	Depressive reaction	17
Cerebral vascular accident	13	Schizophrenia	10
Convulsive disorder	5	Anxiety reaction	10
Cerebral arteriosclerosis	5	Psychosomatic disorders	6
Cortico-striato-spinal disease	2	Character disorder	4
Alzheimer's disease	2	Paranoid state	1
General paresis	1	Conversion reaction	1
Brain tumor	1	Phobic reaction	1
Chronic brain syndrome, alcohol	1		
Total	50	Total	50
Average Age	42.5	Average Age	45.0
Years of formal schooling	9.4	Years of formal schooling	10.0

likely to be of significance in determining aftereffect perception. These variables are the speed at which the spiral is rotated and the length of time for which *S* observes the rotating spiral. In setting up this study an attempt was made to control as many as possible of the factors which have been suggested above as possible explanations for the range of results found by other investigators.

METHOD

Subjects. Three groups of *Ss* were included in this study.

1. The Cortical Damage Group consisted of 48 male and 2 female patients considered by the physicians in charge of their respective cases to have demonstrable organic cortical damage. Many of them, of course, also had damage at subcortical levels. These patients were recommended for this study by their case doctors or other interested staff members, as not too severely aphasic or otherwise distressed to be able to cooperate in the experiment. All of the available information concerning the patient's brain damage was gathered from his clinical chart and from a short conference with his case physician, usually a resident in neurology or neurosurgery. The

diagnoses listed for these patients at the time of testing are given in Table 1.

2. The Neuropsychiatric (NP) Control Group consisted of 42 male and 8 female patients with no known or suspected organic brain pathology, who were referred for routine psychological testing. The older patients in this group (60 years and older) were specifically recommended for this study by their physicians as showing no clinical signs of organic brain damage. Diagnoses of this group are also listed in Table 1. The patient groups were comparable in age and educational level, as indicated in Table 1. All patients in this study had been hospitalized for 6 months or less, although for many of them this was not a first admission.

3. The College Student Group consisted of 35 female and 15 male undergraduates who were required to take part in a number of experiments to fulfill their Introductory Psychology course requirements. This group was, of course, considerably younger and better educated than the patient groups, and unhospitalized.

A total of 2 brain injured and 6 psychiatric patients had to be eliminated from the study because they were unable or unwilling to cooperate.

Apparatus. The spiral used here was a 920-degree Archimedes spiral made from a tracing of the one used by Price and Deabler (1955), altered to a diameter of 7 inches. It was rotated by a device with

a variable speed control which permitted rotation at speeds of from 18 to 90 rpm and a pulley arrangement which allowed rotation in either direction. This apparatus was located slightly below eye level for the average *S* at a viewing distance of approximately 6 feet.

Experimental design. The rotation speeds used were selected on the basis of a preliminary study with eight normal *Ss* as speeds which represented an approximately optimal condition (90 rpm), a definitely nonoptimal condition (18 rpm), and the midpoint between these extremes (54 rpm). Two exposure times were used. Ten seconds appeared to be a near-minimum time for SAE perception, while the 30-second exposure was one which earlier investigators had employed successfully, obtaining aftereffect reports from nearly all of their normal *Ss*.

Ten trials were presented to each *S* in a fixed order of decreasing difficulty as determined by the preliminary study. The first eight trials represented all combinations of 18 and 54 rpm, 10- and 30-second exposure times, and clockwise (CW) and counter-clockwise (CCW) directions, providing a $2 \times 2 \times 2$ factorial design applied to each *S*. The last two trials were run under conditions similar to those employed in most previous clinical studies, thus providing a basis for comparison with the results of earlier investigators.

Procedure. *Ss* were asked if they could see the spiral clearly, and were instructed to look directly at it during all trials. They were told that the spiral at it would be rotated and that they were to watch for apparent expansion, contraction, or changes in depth or distance. The exact wording varied somewhat depending upon *Ss*' educational level. *Ss* were asked to describe what the spiral appeared to be doing while it was rotating, and again when it was halted. Just before stopping the rotation on each trial, the experimenter reminded *S* that he was to keep looking at the spiral even when the apparatus was turned off. All responses were recorded as nearly verbatim as possible, and in doubtful cases a brief inquiry was conducted at the end of the 10 regular trials.

TABLE 2
PERCENTAGE OF SUBJECTS REPORTING AFTERAFFECT
UNDER EACH SET OF CONDITIONS

Conditions	Cortical Damage	NP Control	Student
18-CCW-10	6%	60%	40%
18-CW-10	12	70	74
54-CW-10	16	88	90
54-CCW-10	16	72	60
18-CCW-30	16	78	76
18-CW-30	22	94	94
54-CW-30	42	92	100
54-CCW-30	26	88	86
90-CCW-30	30	84	98
90-CW-30	52	88	100

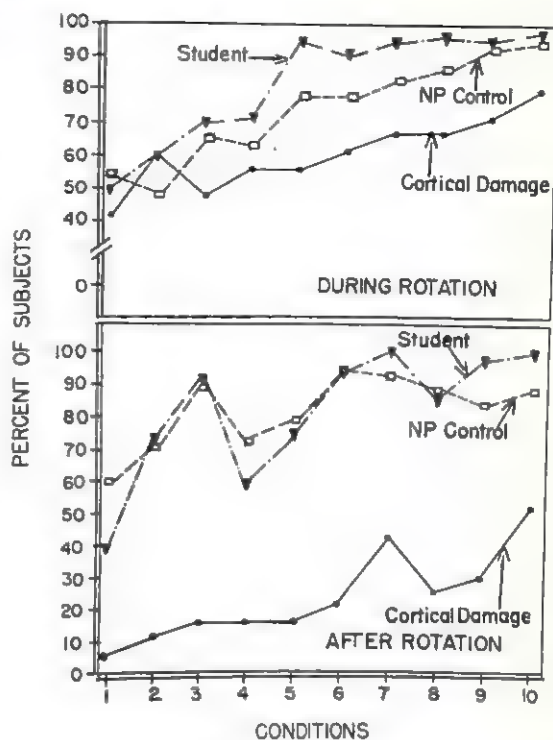


FIG. 1. Percentage of subjects reporting apparent size and/or depth changes.

RESULTS

Of primary interest in this study were the effects of stimulus variation upon report of the spiral aftereffect by the different groups of *Ss*. Table 2 shows the percentage of *Ss* in each group reporting the SAE under each combination of conditions. The sets of conditions are listed in the fixed order in which they were presented, with rotation speed given first, followed by the direction of rotation, and then by the exposure time. Clockwise rotation normally produces an aftereffect of expansion, while the CCW rotation usually produces a contraction aftereffect.

Although the SAE was of particular concern, the responses given both during rotation of the spiral and immediately after its cessation were analyzed. Figure 1 presents the percentage of *Ss* in each group who reported apparent size and/or depth changes during rotation and during the usual aftereffect period for each set of conditions. The sets of conditions are listed in the order in which they were presented. Obviously there is little difference among these groups with regard to

TABLE 3

SUMMARY OF ANALYSES OF VARIANCE OF PROPORTIONS OF SUBJECTS REPORTING AFTEREFFECT

Source	df	Mean Square		F	
		Nonorganic	Organic	Nonorganic	Organic
Exposure Times (A)	1	0.4857	0.2560	48.57**	12.80**
Rotation Speeds (B)	1	0.1567	0.1580	15.67**	7.90*
Rotation Direction (C)	1	0.4349	0.0623	43.49**	3.12
A × B	1	0.0210	0.0070	2.10	0.35
A × C	1	0.0042	0.0099	0.42	0.49
B × C	1	0.0004	0.0001	0.04	0.01
A × B × C	1	0.0101	0.0199	1.01**	1.00
Error		0.0100	0.0200		
Total	7				

* $p < .01$.** $p < .001$.

perception of apparent size and depth changes during rotation of the spiral. A medians chi square test done on these data supports this conclusion. There is, however, a clear difference with regard to the SAE. While the Student and NP Control Groups respond very similarly, the Cortical Damage Group reports the SAE much less frequently. The conditions which apparently produce the greatest differentiation of organic and nonorganic Ss are those of medium difficulty. The conditions most nearly approximating those of previous studies discriminate much less well.

Because the response of the Cortical Damage Group was so clearly different from that of the other two groups, separate analyses of variance of proportions were done, using the

method of Walker and Lev (1953). These analyses are summarized in Table 3. All main effects were significant for the combined Student and NP Control Groups, while for the Cortical Damage Group, exposure time and rotation speed had significant effects, but direction of rotation did not. None of the interactions between variables was statistically significant.

Since the clinical efficacy of this phenomenon as a tool for use in the diagnosis of brain damage has been a major point of contention in the literature, each response of every S was assigned a score of 1 if it indicated the normal SAE, 0 if it did not, following the procedure of Gallese (1956). Thus each S received a score which was equal to the number of normal SAE responses. Figure 2 shows the number of Ss receiving each score. As this figure shows, the Cortical Damage Group differs markedly from the other groups in terms of the number of SAE responses reported. Of the 50 Cortical Damage Ss, 44 reported the SAE fewer than 6 times in 10 trials, while 46 of the 50 NP Controls reported the SAE on 6 or more of their 10 trials. A medians chi square test on these data indicates that this difference between groups is significant well beyond the .001 level.

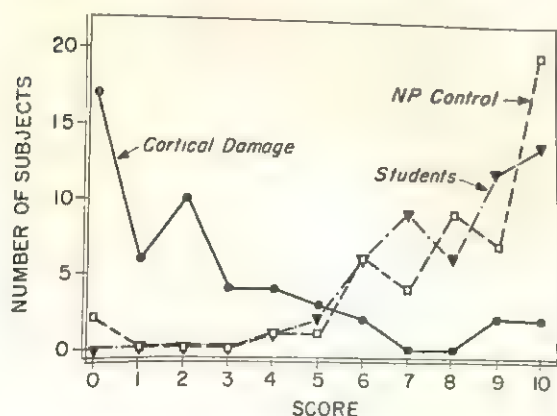


FIG. 2. Number of subjects receiving each SAE score.

Information concerning the location of each patient's brain damage, insofar as it could be ascertained, is presented in Table 4, along

TABLE 4
AFTEREFFECT "SCORES" FOR DIFFERENT
AREAS OF BRAIN DAMAGE

Location of Damage	Mean	SD	Range	N
Generalized, bilateral	1.26	1.33	0-5	19
Generalized, right side	2.00	2.08	0-6	11
Generalized, left side	1.67	—	0-3	3
Bilateral Frontal	1.00	—	0-2	2
Bilateral Frontal-Temporal	0.00	—	—	1
Left Frontal	10.00	—	—	1
Left Frontal-Temporal	5.20	2.58	2-9	5
Left Temporal	1.00	—	—	1
Left Occipital	2.00	—	0-4	2
Right Frontal-Temporal	0.00	—	—	1
Right Temporal	9.00	—	—	1
Right Parietal	4.67	—	2-10	3
Bilateral Damage	1.18	1.23	0-5	22
Unilateral Damage	3.25	3.12	0-10	28

with data on the SAE scores for patients with damage in each area. Only 2 of the 50 patients in this group came to autopsy during the period of this study, and even in these cases complete histological examination of the brain was not carried out. Therefore, it must be recognized that the location of damage is not always clearly delineated or precisely placed. These data, based on the best information available, can only be offered as being suggestive, not conclusive. Unfortunately, the number of patients with damage apparently limited to any one discrete area is too small to permit drawing definite conclusions about the relationship between the location of damage and report of the SAE. There do, however, appear to be differences associated with the extent of damage, bilateral or generalized damage being associated with lower scores than unilateral or localized damage. A *t* test, with correction for heterogeneity of variance, indicated that the unilateral vs. bilateral difference was statistically significant ($p < .01$).

DISCUSSION

As might have been anticipated from the work done on other perceptual phenomena, the results of this study show that both shortening the time of exposure to the rotating spiral and decreasing its rotation speed make perception of the SAE more difficult. This is true for both brain damaged and nonorganic Ss. From the analyses of variance presented in Table 3, it appears that the variable which contributes most to the probability of occurrence of the SAE is the length of time for

which the rotating spiral is observed. The amount of variance attributable to rotation speed is also statistically significant for all groups, but the actual sums of squares are much smaller than those for exposure time. It is interesting that direction of rotation should be a significant variable for the nonorganic Ss but not for the organic group. Apparently this is due to the fact that perception of the SAE was so difficult for the latter group under all but the most optimal conditions that the increment of difficulty added by CCW rotation made very little overall difference.

The results of this study are clearly in agreement with Price and Deabler's (1955) contention that report or nonreport of the SAE discriminates between brain damaged and nonorganic Ss with a fairly high degree of accuracy. Using the series of stimulus conditions developed for this experiment and a cutoff score at the median point for all patients, i.e., between five and six reports of the SAE in 10 trials, 88% of the Cortical Damage and 92% of the NP Control patients are correctly identified. However, it must be pointed out that these Ss are not a random sample from a general hospital population, but rather specific patients for whom there was evidence of organicity or a lack of such evidence.

The patients who are misclassified by this method of scoring can be clearly separated into two categories. Two of the Ss in each group received borderline scores, i.e., reported the SAE sufficiently often to be just above or just below the cutoff point. Four Cortical Damage Ss and two NP Controls are markedly atypical of their respective groups in report of the SAE.

The two NP Control Ss who failed to report the SAE on any trial were both over 60 years of age. While neither showed any obvious clinical signs of deterioration such as are usually associated with brain damage, one could argue that in the absence of histological controls, organic brain pathology in Ss of this age could very well have been undetected in the general clinical picture, while still affecting SAE report. It should be pointed out, however, that many elderly patients received maximal or near-maximal SAE scores in this study.

The four Cortical Damage Ss who reported the SAE 9 or 10 times in 10 trials show a definite similarity in their respective clinical pictures, although the diagnosed area and type of damage is different in each case. Each of these patients has had very mild damage, always very localized (as far as is known), and each shows little or no neurological or intellectual impairment. This would again support the conclusion that size and severity of damage are highly important factors in determining whether or not the SAE is reported. Extensive and/or severe damage to any part of the cortex appears much more likely to destroy the SAE than is localized damage. This may very well be one of the principal reasons for the great variation in the results reported by different investigators who have studied the spiral as a clinical diagnostic device. It seems a likely explanation for the "normal" responses of many of the lobotomy cases and epileptic Ss studied by various researchers.

There is still the possibility, however, that difficulty in verbal report, which might be expected to correlate with the severity of organic damage, is the crucial factor. Both Gallese (1956) and Aaronson (1958) have raised this possibility. Furthermore, Schein (1960) and Van de Castle and Strong (1957) noted that with many patients who did not report the SAE spontaneously, a slight alteration of the test stimulus would elicit such a report. The latter findings could mean that these patients were having difficulty in verbalizing the SAE, much as Aaronson's patients with anomia apparently did. It could also suggest, however, that this was a reflection of psychological rigidity, of the inability to shift one's psychological set, which has so frequently been described as characteristic of the brain injured. Because of the explicit instructions used in the present study and the inquiry conducted in doubtful cases, it is unlikely that the differences in SAE report seen here can be attributed primarily to verbal inefficiency. Furthermore, in the analysis of responses given during rotation of the spiral it was found that all except three of the Cortical Damage Ss reported the occurrence of apparent size and/or depth changes during rotation, indicating that it was not an inability to perceive, conceptualize, or verbalize such

phenomena which led to nonreport of the SAE.

Nevertheless, the results of the above studies raise a very important question. Can it be said that many brain damaged patients do not perceive the SAE, i.e., do not consciously experience it unless a new test figure is substituted for the original? If so, this implies that the apparent loss of the SAE in brain damaged patients is due to something more than just a destruction of sensory elements in the cortical visual system. It would appear that for spontaneous perception of the SAE not only must the neuronal chains from retina to visual cortex be reasonably intact, but also that other parts of the brain which are involved in organizing and transforming sensory information into conscious awareness must be functional. If this is correct, it should not be surprising that mild, focal damage almost anywhere in the cortex does not interfere with the SAE, while larger, more severe damage usually does.

It appears to me that much of our current theorizing concerning the apparent destruction of the SAE by brain damage has been much too narrow in scope. We have been attempting to seek out a single comprehensive factor which would explain both the failure of many brain damaged Ss to report the SAE and the contradictory evidence which has been reported in several studies. The factor of impaired verbal report is perhaps the most frequent "cause" invoked to explain this discrepancy. I believe that in doing so, we have been very much inclined to oversimplify what is in actuality an extremely complex process. Because of the striking character of the phenomenon, it is easy to forget that the perception of the SAE is in itself a highly complex operation. Multitudinous factors are involved in determining whether or not a particular S will perceive the SAE on any given trial. In the present study many Ss, even among the college students, sometimes failed to perceive the SAE under what are usually favorable conditions after having reported the phenomenon under more difficult conditions. This study has provided a systematic investigation of several stimulus variables, with certain subject variables controlled, but there are many variables which were not, and perhaps

could not be controlled in such a clinical experiment. We know from the work of Wohlge-muth (1911) and others that such variables as the size and viewing distance of the spiral, the state of light or dark adaptation of the eye, the intertrial interval, the constancy of fixation, and the fatigue state of the *S* can all be important. Brain damage, then, is not the only factor which may prevent the occurrence of the SAE, and likewise, brain damage alone may not prevent such occurrence under otherwise favorable conditions.

Not only is the SAE a complex phenomenon, but the brain itself is such an extremely complex structure with at least some degree of localization of function, that it seems to me completely unreasonable to expect that brain damage as a general clinical entity should affect SAE perception in the same way in each patient. Damage of different degrees of severity to different areas of the brain will produce different effects, any one or more of which may be crucial with regard to SAE report in a given case. Frontal lobe damage, especially with involvement of the prefrontal region, frequently interferes with the ability of *Ss* to attend or concentrate and greatly increases their distractibility (Peale, 1954). Thus, in many cases the SAE may not be perceived primarily because of the inability to maintain fixation. Goldstein (1942) tells us also that damage to the frontal lobes impairs the patient's ability to take and hold the abstract attitude. His methods for the measurement of such impairment would suggest that this is closely related to an inability to shift psychological set, and it may be that such patients cannot consciously experience the SAE unless something is done to break their original set, as Schein (1960) did. In still other cases, damage to the temporal lobe producing an aphasic disturbance could result in an inability to report the SAE, even if it is consciously perceived. Both Teuber and Bender (1949) and Werner and Thuma (1942) found that unusually short interstimulus intervals were necessary to produce apparent movement phenomena in their brain damaged *Ss*, suggesting that some sort of psychophysiological deficit was occurring in the integration of discrete visual stimuli. In many cases with diffuse and severe cortical damage

the cortical visual system and/or "association" cortex may be so disrupted as to prevent reception and integration of the stimuli impinging upon the retina.

In the present study, the use of some stimulus conditions which make SAE perception difficult even for normal *Ss* added a further increment of difficulty to those noted above. Thus, in addition to the different physiological effects produced by shorter, more slowly moving stimuli, the unusually long series of trials increased the tendency for physical and psychological fatigue to interfere with *Ss*'s concentration on the task at hand. These additional handicaps for the brain damaged patient, whose SAE report may already have been impaired by one or more of the factors suggested above, probably produced the relatively clear-cut differentiation of brain damaged and nonorganic groups in the present study.

SUMMARY

This study was designed to investigate the effects of certain stimulus variables upon perception of the spiral aftereffect (SAE). A 920° Archimedes spiral was presented to 50 patients with cortical brain damage, 50 psychiatric patients with no known or suspected brain damage, and 50 college students in a standard series of trials. These trials contained all combinations of 18 and 54 rpm rotation speeds, 10- and 30-second exposure times, and clockwise and counterclockwise rotations. One additional trial in each direction was given at 90 rpm with a 30-second exposure time, approximating the conditions used by other investigators.

It was found that during rotation of the spiral the patient groups reported apparent size and/or depth changes with approximately equal frequency. Following rotation, however, 46 of the 50 NP Control *Ss* reported the SAE 6 or more times in 10 trials, while only 6 of the Cortical Damage *Ss* did likewise. The college students responded almost identically with the NP Controls. The atypical psychiatric patients tended to be elderly people, raising the possibility of undiagnosed brain damage associated with aging, while the atypical Cortical Damage *Ss* were characterized by mild and localized damage, with little or no

clinical neurological or intellectual impairment. Although histological evidence was not available, the data strongly suggest that location of damage is relatively unimportant, but that extensive or severe damage to any part of the cortex markedly reduces the probability of SAE report.

Analyses of variance indicated that exposure time and rotation speed were significant factors in perception of the SAE for all subject groups, while direction of rotation was a significant variable only for the nonorganic groups. Conditions of medium difficulty discriminated best between the patient groups, while conditions similar to those of previous studies were less discriminating.

It was concluded that multiple factors are involved in the disruption of SAE report in brain damaged Ss, and that the increment of difficulty added by a long series of relatively difficult trials contributed to the clear-cut discrimination between groups in this study.

REFERENCES

- AARONSON, B. S. Age, intelligence, aphasia and the spiral aftereffect in an epileptic population. *J. clin. Psychol.*, 1958, 14, 18-21.
- DAVIDS, A., GOLDENBERG, L., & LAUFER, M. W. The relation of the Archimedes spiral aftereffect and the Trail Making Test to brain damage in children. *J. consult. Psychol.*, 1957, 21, 429-433.
- GALLESE, A. J. Spiral aftereffect as a test for organic brain damage. *J. clin. Psychol.*, 1956, 12, 254-258.
- GARRETT, E. S., PRICE, A. C., & DEABLER, H. L. Diagnostic testing for cortical brain impairment. *AMA Arch. Neurol. Psychiat.*, 1957, 77, 223-225.
- GILBERSTADT, H., SCHEIN, J. D., & ROSEN, A. Further evaluation of the Archimedes spiral aftereffect. *J. consult. Psychol.*, 1958, 22, 243-248.
- GOLDSTEIN, K. *Aftereffects of brain injuries in war*. New York: Grune & Stratton, 1942.
- HAMMER, E. R. Temporal factors in figural aftereffects. *Amer. J. Psychol.*, 1949, 62, 337-354.
- PAGE, H. A., RAKITA, G., KAPLAN, H. K., & SMITH, N. B. Another application of the spiral aftereffect in the determination of brain damage. *J. consult. Psychol.*, 1957, 21, 89-91.
- PEALE, T. L. *The neuroanatomical basis for clinical neurology*. New York: McGraw-Hill, 1954.
- PHILBRICK, E. B. The validity of the spiral aftereffect as a clinical tool for diagnosis of organic brain pathology. *J. consult. Psychol.*, 1959, 23, 39-43.
- PRICE, A. C., & DEABLER, H. L. Diagnosis of organicity by means of spiral aftereffect. *J. consult. Psychol.*, 1955, 19, 299-302.
- SCHEIN, J. D. Personal communication, 1960.
- SPIVAK, G., & LEVINE, M. The spiral aftereffect and reversible figures as measures of brain damage and memory. *J. Pers.*, 1957, 25, 767-777.
- TEUBER, H. L., & BENDER, M. Alterations in pattern vision following trauma in the occipital lobe in man. *J. gen. Psychol.*, 1949, 40, 37-57.
- VAN DE CASTLE, R. L., & STRONG, P. N. A further study on the diagnostic efficiency of the spiral aftereffect. Paper read at Southeastern Psychological Association, Nashville, Tennessee, March 1957.
- WALKER, H. M., & LEV, J. *Statistical inference*. New York: Holt, 1953.
- WERNER, H., & THUMA, A. A deficiency in the perception of apparent motion in children with brain injury. *Amer. J. Psychol.*, 1942, 55, 58-67.
- WOHLGEMUTH, A. On the after-effect of seen movement. *Brit. J. Psychol.*, 1911, Monogr. Suppl. 1, 1-117.

(Received February 25, 1960)

INTELLECTUAL FUNCTIONING IN A GROUP OF NORMAL OCTOGENARIANS¹

ANDREW S. DIBNER AND JAMES F. CUMMINS

Veterans Administration Outpatient Clinic, Boston, Massachusetts

This paper presents some empirical findings on the intellectual functioning of 50 men who were in service during the Spanish American War and who now average 80 years of age. These men live in the Greater Boston Area and responded to an invitation to attend the research oriented Geriatric Clinic established at the Boston Veterans Administration Outpatient Clinic in 1958. The functions of the clinic as well as social data on these men are described elsewhere by Nichols and Cummins (in press).

The questions asked by this study were the following: First of all, how would these 80-year-old, relatively healthy men perform on the Wechsler-Bellevue Intelligence Scale (Wechsler, 1944) and how would their results compare with those reported in other studies of older people? Secondly, is there evidence of intellectual decline as measured by this test, and if so, how does this evidence compare with findings on other age groups; and, is this intellectual decline related to level of intelligence as has been sometimes claimed? Thirdly, do the time limits established for the Wechsler-Bellevue scale, Form I, inordinately penalize the slower, older person?

The Wechsler-Bellevue Form I scale was employed rather than the newer Wechsler Adult Intelligence Scale (Wechsler, 1953) because there are more data available on Form I performance by various pathological groups and normal age groups with which these data might be compared.

¹ Based upon a paper presented at the twelfth Annual Meeting of the Gerontological Society, Inc., Detroit, Michigan, November 1959, resulting from an investigation carried out at the Boston Veterans Administration Outpatient Clinic.

METHOD

Subjects. Table 1 presents some descriptive information about the subjects. The Total Group of 50 men range in age from 73 to 89 years, with mean age at 80.4, standard deviation 2.1 years. Their education ranges from no formal education to college graduate (16 years). Mean education for the Total Group is 7.9 years. Their highest achieved occupational status as rated according to Warner's scale (Warner, Meeker, & Eels, 1949) ranges from 1 (the highest level, as professional) to 7 (the lowest level, as laborer) with a mean occupational level of 4.2.

For the purpose of intragroup analysis the Total Group was divided into Younger and Older Groups. The 23 subjects in the Younger Group range in age from 73 to 79 years and their mean age is 77.6. The men in the Older Group range in age from 80 to 89 and average 82.8 years, approximately 5 years older than the men in the Younger Group. As can be seen from Table 1, the two groups are highly similar in education and highest achieved occupational level.

Procedure. Subjects were administered the 11 subtests of the Wechsler-Bellevue Intelligence Scale, Form I, by experienced examiners using standard instructions with the exception of extending time limits on all timed tests except Digit Symbol. The method is similar to that employed by Doppelt and Wallace (1955) and was done in order to compare performance under standard (ST) and extended time (ET) limits.

Although the Vocabulary subtest was administered it was not considered in determining the Verbal weighted scores or IQs in order to make these results more comparable with those of other investigators. The IQs were determined by the extrapolation technique suggested by Wechsler (1944) for age groups beyond those given in his tables.

RESULTS

Table 2 presents the Verbal, Performance, and Total weighted scores and IQs of the Younger, Older, and Total Groups. With respect to weighted scores, differences between groups are slight and nonsignificant; though curiously, the Older subjects outdo the Younger on the performance scale, where they

TABLE 1
AGE, EDUCATION, AND OCCUPATION OF SUBJECTS

Group	N	Age in Years			Education in Years			Occupational Rating (Warner's Scale)		
		Range	M	SD	Range	M	SD	Range	M	SD
Younger	23	73-79	77.6	1.6	4-14	8.0	2.2	1-6	4.3	1.3
Older	27	80-89	82.8	2.6	0-16	7.8	3.5	1-7	4.0	1.5
Total	50	73-89	80.4	2.1	0-16	7.9	2.9	1-7	4.2	1.4

TABLE 2

VERBAL, PERFORMANCE, AND FULL SCALE WEIGHTED SCORES AND IQs FOR YOUNGER, OLDER, AND TOTAL GROUPS

Scores	Younger Group		Older Group		Total Group	
	M	SD	M	SD	M	SD ₁
Verbal WS	47.2	7.3	44.1	11.8	45.3	10.1
Performance WS	32.6	9.4	33.1	9.6	32.8	9.5
Full Scale WS	79.1	14.3	78.0	18.1	78.5	16.3
Verbal IQ	110.5	6.9	109.2	10.8	109.8	9.2
Performance IQ	110.5	10.0	113.7	9.9	112.1	10.1
Full Scale IQ	108.7	8.2	109.9	10.2	109.3	9.3

might be expected to do worse. This result is probably a sampling fluctuation. The superiority of verbal weighted score over performance weighted score is statistically significant for both Older and Younger subjects.

With respect to IQ, there are no significant differences between groups. The mean IQ, using Wechsler's extrapolation method, is 109.

Table 3 presents the means and standard deviations of subtest scores for the subgroups and Total Group. The highest subtest scores for the Total Group are Vocabulary, Information, and Comprehension, with Average weighted scores of 10 or 11. The next lower group of scores, averaging about 8 or 9, were those of Arithmetic, Similarities, Picture Completion, and Object Assembly. The lowest grouping of scores were those on Digit Span, Block Design, Picture Arrangement, and Digit Symbol, averaging 5 or 6.

It is worthy of note that the Information, Comprehension, and Vocabulary scores above 10 are consistent with the above average IQs yielded for the test as a whole by Wechsler's extrapolation method.

One may examine the pattern of differential decline of the various subtests by two means: first by comparing the differential scatter of subtests which occurred within the two age groups, and then by comparing these subjects' scatter of subtest scores with that reported by other investigators with aged subjects.

Figure 1 demonstrates the similarity of subtest scatter for the Younger and Older Groups within the sample. First, we note that the Younger men score higher than the Older men on 8 of the 11 subtests, although none of these differences is significant by *t* test. Secondly, we note that the pattern of performance of the two groups on the various subtests is highly similar. This is evidence within this particular sample of the reliability of the differential decline of subtest performance which has been reported for age by Wechsler (1953) and others.

TABLE 3
MEANS AND STANDARD DEVIATIONS OF SUBTEST WEIGHTED SCORES FOR YOUNGER, OLDER, AND TOTAL GROUPS

Subtest	Younger Group		Older Group		Total Group	
	M	SD	M	SD	M	SD
Information	11.4	1.5	10.6	2.7	11.0	2.3
Comprehension	10.6	2.1	10.2	3.0	10.4	2.6
Digit Span	7.2	2.9	6.4	2.6	6.7	2.6
Arithmetic	9.2	2.5	8.6	2.9	8.9	2.7
Similarities	8.8	2.3	8.1	3.5	8.4	2.9
Vocabulary	11.8	2.5	11.0	3.6	11.3	1.9
Pic. Arrangement	6.2	3.1	5.0	2.5	5.5	2.8
Pic. Completion	8.1	3.1	9.4	3.2	8.7	3.2
Block Design	5.8	1.8	5.6	2.5	5.7	2.2
Obj. Assembly	7.7	3.0	8.6	2.4	8.1	2.7
Digit Symbol	4.9	1.7	5.0	1.7	4.9	1.7

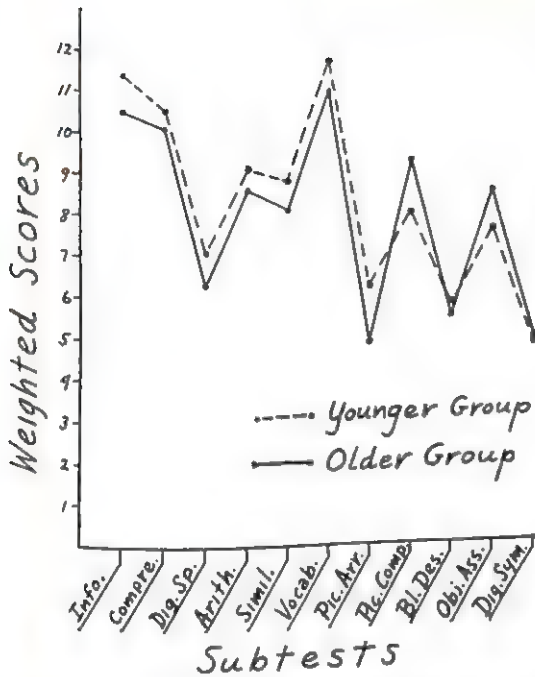


FIG. 1. Mean weighted subtest scores for younger and older groups.

Table 4 presents the subtest performance of this aged group contrasted with that of others reported in the literature, with subtests ranked in terms of their average weighted scores. Although these various samples were drawn partly from institutions and partly from the community, there seems to be good agreement at the ends of the scale. As indicated by the averages of the ranks across

studies, the abilities contributing to performance in the Information, Comprehension, and Arithmetic subtests are best retained in the older years, and those involved in performance on the Block Design, Picture Arrangement, and Digit Symbol subtests are least well retained. The major factor involved in the last mentioned three is probably speed.

The four subtests ranked in the middle range do not clearly distinguish themselves from each other in terms of differential decline.

Wechsler developed a method of estimating deterioration of functioning by comparing the sum of scores of the subtests which have been found to decline more slowly with age with the sum of scores of those subtests on which performance tends to be significantly impaired with age. This so-called deterioration was measured in the present group and compared with Wechsler's norms for younger age groups. The result is shown in Figure 2.

Using Wechsler's formula for deterioration quotient (or DQ) shown in the lower left-hand corner of Figure 2, the average DQ for the Younger Group was found to be 70.1 and for the Older Group to be 66.5. These are plotted along with the DQs for various age groups reported by Wechsler. They seem to fall directly on an extrapolation of the empirical curve determined by Wechsler. Thus, the same rate of decline of abilities relative to each other is occurring in this sample as in

TABLE 4

SUBTESTS RANKED BY AVERAGE WEIGHTED SCORE IN PRESENT STUDY AND OTHERS

[illegible]

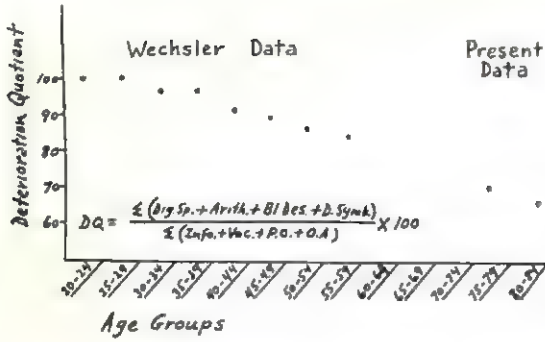


FIG. 2. Average deterioration quotients at different ages—Wechsler's norms and present results.

Wechsler's sample despite the fact that the present sample is of higher average intelligence. The DQ, as an intratest ratio for each individual, is relatively independent of the level of subtest scores.

The relation of deterioration to present level of intelligence was tested further by correlating DQ with IQ. The correlation was .03. DQ was then correlated with two other measures which can be considered to be related to past functional intelligence, namely, education and highest occupational level reached. The correlations were $-.01$ and $.01$. Thus no evidence was found of relationship between present or past intellectual level and amount of present intellectual decline.

Now let us consider the effects of using standard time limits or extended time limits with this older group.

Table 5 presents the changes in weighted scores on four timed subtests which occurred by allowing subjects extra time. Twelve percent of the subjects increased their Arithmetic score an average of 1.5 points, resulting in a rise in the group average of only 0.2 of a

TABLE 5
MEAN WEIGHTED SCORES AND CHANGES IN SCORES ON
FOUR SUBTESTS USING STANDARD AND
EXTENDED TIME LIMITS

	Arithmetic		Picture Arrangement		Block Design		Object Assembly	
	ST	ET	ST	ET	ST	ET	ST	ET
Mean	8.9	9.1	5.5	5.8	5.7	6.6	8.1	8.6
SD	2.7	2.7	2.8	2.8	2.2	2.5	2.7	2.3
Percentage of Group Changed	12%		16%		43%		18%	
Average Change	1.5		1.6		2.1		2.8	
rho	.96		.96		.89		.89	

point. The change in the Picture Arrangement subtest was also slight, as was that in the Object Assembly subtest. However, 43% of the subjects improved their Block Design score when given extra time. These changes did not appreciably disturb the rank-ordering of the subjects within the sample as shown by the high correlations between scores under standard and extended conditions.

Table 6 presents the changes in IQ which resulted from using extended time limits. Only 12% of the subjects increased their Verbal IQ, and the group average raised only slightly from 109.8 to 110.1, however, 60% of the Performance IQs were raised, bringing the average for the group up from 112 to almost 114, and 61% of the Full Scale IQs were raised by allowing extra time, however raising the group average but one point. The rank correlations between IQ scores under standard and extended time conditions were extremely high, indicating that these older subjects do not appreciably change their ranking in relation to each other as a result of having extended time limits.

TABLE 6
MEAN VERBAL, PERFORMANCE, AND FULL SCALE IQS AND CHANGES IN IQS
USING STANDARD AND EXTENDED TIME LIMITS

	Verbal IQ		Performance IQ		Full Scale IQ	
	ST	ET	ST	ET	ST	ET
Mean	109.8	110.1	112.1	113.9	109.3	110.3
SD	9.2	9.2	10.1	9.8	9.3	9.3
Percentage of Group Changed	12%		60%		61%	
Average Change	1.5		3.2		1.7	
rho	.98		.95		.99	

SUMMARY

A group of 50 relatively healthy, male veterans of the Spanish American War, who now average 80 years of age, was found to be above average in intelligence, performing well in tests measuring retention and comprehension of verbal material but performing poorly in tests affected by psychomotor speed and abstract thinking. The pattern of decline of various abilities tested is consistent within the younger and older men of the sample and consistent with other studies of older groups. The deterioration quotient for the Total Group follows closely an extrapolation of the curve of deterioration with age empirically derived by Wechsler. Extent of intellectual decline was found to be unrelated to level of intelligence as measured by IQ, education, or highest occupational level reached.

The use of standard time limits was found to appreciably affect the older person's score on the Block Design subtest, but not the Arithmetic, Picture Arrangement, or Object Assembly subtests. Standard time limits were found to depress the Performance IQ two points and Full Scale IQ one point, but do not appreciably affect older persons' relative rankings within their group as far as IQ scores are concerned.

REFERENCES

- CHESROW, E. J., WOSIKA, P. H., & REINITZ, A. H. A psychometric evaluation of aged white males. *Geriatrics*, 1949, 4, 169-177.
- DOPPELT, J., & WALLACE, W. Standardization of the WAIS for older persons. *J. abnorm. soc. Psychol.*, 1955, 51, 312-330.
- FOX, C., & BIRREN, J. E. The differential decline of subtest scores on the Wechsler-Bellevue Intelligence Scale in 60-69-year-old individuals. *J. genet. Psychol.*, 1950, 77, 313-317.
- HOWELL, R. J. Changes in Wechsler subtest scores with age. *J. consult. Psychol.*, 1955, 19, 47-50.
- MADONICK, M. J., & SOLOMON, M. The Wechsler-Bellevue scale in individuals past sixty. *Geriatrics*, 1947, 2, 34-40.
- NICHOLS, M. R., & CUMMINS, J. F. The lifelong social adjustment of a group of normal octogenarians. *Geriatrics*, in press.
- RABIN, A. I. Psychometric trends in senility and psychoses of the senium. *J. gen. Psychol.*, 1945, 32, 149-162.
- WARNER, W. L., MEEKER, M., & EELS, K. *Social class in America*. Chicago: Science Research Associates, 1949.
- WECHSLER, D. *Measurement of adult intelligence*. (3rd ed.) Baltimore: Williams & Wilkins, 1944.
- WECHSLER, D. *The measurement and appraisal of adult intelligence*. (4th ed.) Baltimore: Williams & Wilkins, 1953.

(Received March 10, 1960)

PREDICTION OF RELAPSE FOR PSYCHIATRIC PATIENTS

DAVID J. GOUWS¹

Western Psychiatric Institute and Clinic

In a review of the literature on prognostic measures in psychopathology Windle (1952) has commented favorably on Feldman's (1951) *Ps* scale. This scale was developed to predict the probability of success of shock treatment and consists of 52 items derived from the Minnesota Multiphasic Personality Inventory. Windle (1952) suggested that "further work in prognosis should employ this scale, or just as desirable, previously gathered data should be reanalyzed with it" (p. 466). Only two reports on cross-validation studies of the *Ps* scale were located (Pumroy & Kogan, 1958; Roberts, 1959). Both failed to confirm Feldman's own cross-validation findings. This paper, which reports a reanalysis of previously gathered data, presents further validation evidence for, and suggests possible interpretations of, the *Ps* scale.

SAMPLE DESCRIPTION

A search of the files yielded complete MMPI records for 104 inpatients of the Western Psychiatric Institute and Clinic, tested between 1944 and 1951, but mostly in the period 1946-47. Of these, 6 were excluded on the basis of major organic involvement, and 4 because they had obtained a "?" raw score in excess of 100. This left 94 cases with usable MMPI records.

As the reasons for the original test referrals were in many cases unknown, these 94 patients could not be regarded as an unselected sample of the hospital population at that time. However, since any bias in this sample presumably would not affect the range of variation of either the predictors or the criterion used, the possible unrepresentativeness of the sample was not regarded as a limitation.

Feldman (1951) listed several criteria which have to be met before his *Ps* scale can be validly employed. Of the 94 cases, 60 met all these requirements, while the other 34 failed to meet only the

requirement of elevated scores on one or more of the critical scales. These groups will subsequently be referred to as the "elevated" and "unelevated" groups, respectively.

The following dichotomous "criterion of improvement" was used in this study: "Improved" cases were all those who obtained an "improved" rating on leaving hospital and who were not readmitted for psychiatric reasons to this or to another hospital in the 5 years following discharge. "Unimproved" cases were those rated "unimproved" on leaving hospital (usually for transfer to another institution), as well as those cases with an improved rating upon leaving hospital, but with a record of a subsequent relapse severe enough to require psychiatric rehospitalization in the 5 years following discharge. (The shortest period of such rehospitalization was 3 months). The information on rehospitalization of former patients had been gathered at 6-month intervals from the referring hospitals and physicians, and is complete, so far as is known, except for some patients who may have been hospitalized in another state without mentioning their previous hospitalization here.

Of the 60 patients in the elevated group 23 had received either or both insulin and electric shock treatment. This treatment started from one day to several months after the administration of the MMPI. The 11 patients subsequently classified as improved in terms of the criterion used received an average of 11 electrically induced convulsions and 8 insulin comas each. The corresponding figures for the unimproved group of 12 patients were 14 and 16. The remaining 37 patients in the elevated group received general supportive hospital care, which often included an appreciable amount of individual psychotherapy. The composition of the various subgroups in terms of diagnosis, sex, and age, is shown in Table 1.

PROCEDURE AND RESULTS

Attempting to determine what is measured by the *Ps* scale, Feldman (1951) has pointed out that an unusually large number of items which refer to interpersonal relationships are included in his *Ps* scale, and that his "recovered" group, although consisting almost entirely of psychotic patients, obtained scores very similar to that of a group of normal subjects. Since quality of inter-

¹ On leave from the University of Pretoria, South Africa. It is a pleasure to acknowledge the help of Edith Fleming in obtaining the follow-up data on the patients used in this study.

TABLE 1
COMPOSITION OF SAMPLE

Subgroup	Age			Diagnosis and Sex					
				Affective Disorder		Schizophrenia ^a		Nonpsychosis ^b	
	N	Mean	SD	M	F	M	F	M	F
Elevated, shock treatment:									
Improved	11	35.0	8.8	1	3	1	1	3	2
Unimproved	12	35.2	7.0	2	0	5	2	1	2
Elevated, supportive care:									
Improved	20	33.4	11.3	1	1	2	3	4	9
Unimproved	17	29.9	10.9	1	3	1	3	3	6
Unelevated, all treatments:									
Improved	19	36.3	10.7	3	2	2	3	5	4
Unimproved	15	35.5	10.1	3	2	1	4	3	2

^a Including two cases of undiagnosed psychoses.^b Psychoneuroses and personality disorders.

personal relationships enters into most concepts of adjustment, it may be asked whether an "adjustment" questionnaire might not differentiate equally well between patients that improve and those that do not. Furthermore, since 38 out of the 52 items comprising the *Ps* scale tended to be answered "True" by Feldman's unimproved criterion group, there may also be a question about the role of response set, specifically of "acquiescence," and about the possibility of using an acquiescence measure as a suppressor variable in predicting improvement. As scores on a 142-item Adjustment key for the MMPI (Fulkerson, 1957) as well as on a 24-item Acquiescence key, also derived from the MMPI (Fulkerson, 1958) were available for the patients in this sample, these questions could be followed up at small extra cost.

As a first step, the total elevated group was split into improved and unimproved cases (according to the criterion described) and the *Ps*, Adjustment, and Acquiescence scores were compared. The results in Table 2 show significant differences between improved and unimproved patients in the case of all three variables.

Although the different follow-up criteria used precluded a strict comparison of the

Ps distributions obtained in the present investigation with those obtained by Feldman (1951), some interesting similarities did appear. The mean *Ps* score of the improved group in this study is about halfway between the means reported for Feldman's recovered and improved test groups, while the present

TABLE 2
QUESTIONNAIRE SCORES OF THE TOTAL
ELEVATED GROUP

Criterion Status	Questionnaire Score				
	N	Mean	SD	<i>t</i>	<i>r</i> _{bis}
<i>Ps</i>					
Improved	31	21.2	8.4	2.68**	.41
Unimproved	29	27.2	8.9		
Adjustment					
Improved	31	55.9	17.8	2.47*	.38
Unimproved	29	66.8	16.4		
Acquiescence					
Improved	31	11.5	3.1	2.73*	.42
Unimproved	29	14.0	3.9		

* $p < .01$.** $p < .005$ (one-tailed *t* tests in the case of *Ps* and Adjustment).

unimproved group obtained a slightly lower (more favorable) mean score than his unimproved group. The overlap between the *Ps* score distributions of the improved and unimproved groups in the present study is somewhat greater than that found by Feldman. Inspection reveals that a cutting score of 24.5 will maximize the total correct predictions of the criterion in the present sample. The analogous cutting scores for Feldman's distributions (maximizing total correct prediction of his criterion) range between 20.5 and 28.5, depending on how the dichotomy is obtained.

The intercorrelations between the *Ps*, Adjustment and Acquiescence scores of the 60 patients in the elevated group were: *Ps*-Adjustment, $r = .84$; *Ps*-Acquiescence, $r = .65$; and Adjustment-Acquiescence, $r = .53$.

The *Ps*-Adjustment correlation coefficient is of the same order as the corrected split-half reliability coefficient of .86 reported for the *Ps* by Feldman (1951). With only 12 items common to these two scales, which were derived under widely different conditions, such a high correlation is remarkable. The high acquiescence loadings (if we assume for the moment the validity of the Acquiescence scale used) of both the *Ps* and the Adjustment scales are according to expectation. Unfortunately, the fact that the Acquiescence scale discriminates as well as do the other two between the improved and unimproved patients, rules out the possibility of using it as a suppressor variable.

It remained to be seen whether prediction of relapse was equally effective for the group who had received general supportive hospital care as for the patients who had received insulin and/or electric shock treatment. Splitting the elevated group into supportive care and shock treatment subgroups, it was found that the *Ps* scale discriminated equally well ($p < .025$) between improved and unimproved patients in both treatment subgroups. This suggests that what the *Ps* scale measures is not so much ability to benefit from shock treatment as the tendency to get well irrespective of type of treatment. Incidentally, the shock treatment subgroup obtained a slightly lower (more favorable)

mean *Ps* score ($.1 > p > .05$) than the supportive care subgroup.

A practical limitation of Feldman's scale is that it can only be used with patients whose MMPI profiles meet the stated requirements. In the total sample of 94 patients available for this study, 34 (or 36%), are thus excluded from consideration. It seemed worthwhile to investigate whether the relationship observed between *Ps* scores and improvement in the elevated group would not be found in the unelevated group as well. Comparison of the *Ps* scores of the improved and unimproved patients in the unelevated subgroup yielded no significant difference, however, neither did the Adjustment or Acquiescence scores differ significantly, although all three differences were in the same direction as for the elevated group. To estimate the influence of truncation of scores on the validity of the *Ps* scale, biserial r was calculated and corrected for restriction of range. The r obtained was .22 and it rose to .31 ($p > .05$) after correction.

Feldman has speculated about the type of patient who is ill enough to be in a psychiatric hospital, yet responds essentially like a normal person to psychological questionnaires. Two incidental but interesting observations on the unelevated subgroup in this study should be reported: Of 8 patients diagnosed as "psychopathic personality" or "psychopathic state" in the original sample of 104 cases, 5 were in the unelevated subgroup. Secondly, 12 patients out of the 34 in the unelevated subgroup attended college, and 8 of them graduated. The corresponding numbers for the elevated subgroup ($N = 60$), were 9 and 2. These differences, tested by χ^2 , with Yates' correction, were significant at the .05 and .01 level for college attendance and graduation, respectively.

DISCUSSION

Although a significant difference between improved and unimproved patients was found, the differentiation in terms of *Ps* scores was not marked enough to enable reliable prediction for the individual patient to be made except in a small minority of

extreme cases. In evaluating the less impressive discrimination obtained—as compared with the data reported by Feldman (1951)—the difference between the improvement criteria used in the two studies, as well as the long period that elapsed between taking the MMPI and starting shock therapy in the case of some patients in this study, should be taken into account. Whether this latter point made any real difference is questionable, as the prediction was shown to hold irrespective of treatment received. It does suggest that the attribute(s) tapped are reasonably stable in time.

Feldman (1951), discussing his own cross-validation findings, has suggested that the *Ps* scale measures "propensity to improve" irrespective of diagnosis or method of treatment. The present findings, that the *Ps* scale predicts improvement equally well for shock treatment and for supportive care patients, and that an Adjustment scale, developed for a different purpose, can predict as well as the *Ps* scale, do seem to confirm the notion that a general characteristic or group of characteristics, rather than a specific characteristic, namely, responsiveness to shock treatment, is being measured by these scales. That an acquiescence measure, the items of which were chosen so as not to discriminate between well and poorly adjusted military personnel, predicts improvement so well, indicates that what Feldman has tentatively labeled "propensity to improve" may be a complex entity.

SUMMARY

In a cross-validation study Feldman's *Ps* MMPI scale—for the prediction of response to shock treatment—was found to discriminate significantly between patients who had, and did not have, a relapse within 5 years, irrespective of treatment. An Adjustment scale, developed for military personnel, differentiated equally well between the criterion groups. These two scales intercorrelated .84, and, respectively, correlated .65 and .53 with an Acquiescence scale. The possible use of the Acquiescence scale as a suppressor variable was explored and the implications of these data for prognosis of psychiatric patients discussed.

REFERENCES

- FELDMAN, M. J. A prognostic scale for shock therapy. *Psychol. Monogr.*, 1951, 65(10, Whole No. 327).
- FULKERSON, S. C. Adaptability screening of flying personnel: Research on the Minnesota Multiphasic Personality Inventory. *USAF Sch. Aviat. Med. Rep.*, 1957, No. 57-106.
- FULKERSON, S. C. An acquiescence key for the MMPI. *USAF Sch. Aviat. Med. Rep.*, 1958, No. 58-71.
- PUMROY, D. K., & KOGAN, W. S. A validation of measures that predict the efficacy of shock therapy. *J. clin. Psychol.*, 1958, 14, 46-47.
- ROBERTS, J. M. Prognostic factors in electroshock treatment of depressive states: II. The application of specific tests. *J. ment. Sci.*, 1959, 105, 703-713.
- WINDLE, C. Psychological tests in psychopathological prognosis. *Psychol. Bull.*, 1952, 49, 451-482.

(Received March 11, 1960)

THE REPRESENTATION OF PHYSIQUE IN CHILDREN'S FIGURE DRAWINGS

A. B. SILVERSTEIN¹

Pacific State Hospital

AND

H. A. ROBINSON

Psychiatric Institute, University of Maryland

The interpretation of human figure drawings as a projective technique is said to rest on the assumption that the drawn figure represents the subject's body image—"the picture of his body which he forms in his mind" (Schilder, 1950). Much of the research purporting to test the "body image" hypothesis has made use of physically disabled subjects; investigations of this kind were reviewed in a previous paper (Silverstein & Robinson, 1956). A search of the literature has revealed but little work based on subjects within the normal range of physical variation. Berman & Laffal (1953) reported that the predominant somatotype of drawn figures was related to that of the men who drew them;² and Kotkov and Goodman (1953) found that figures drawn by obese women tended to cover a greater area of the page than those drawn by women of ideal weight.

In a review of empirical evidence on figure drawings, Swensen (1957) questioned the treatment and interpretation of the data of both of these studies, but even if their seemingly positive results are accepted at face value, it should be noted that neither study provided a direct test of the body image hypothesis. The same is true of investigations of the figure drawings of the physically handicapped. When previous research is considered from an operational viewpoint, it is immediately apparent that its focus has been the relation between the drawn figure

and the actual structure of the body, *not* the body image. Subjects have been selected not on the basis of differences in body image, but because they differed with respect to actual physique.³ While it may be true that normally there is no discrepancy between the body image and the body structure there seems to be no empirical evidence at present to support this common assumption. To the extent that the subject's "mental picture" of his physique does not correspond to his actual physique, the relations and differences observed in previous research on the body image hypothesis are clearly in error.

To the writers' knowledge, the study reported here is the first to distinguish operationally between body image and body structure. The method made it possible to assess the degree of correspondence between body image and body structure, and to perform a direct test of the body image hypothesis, i.e., to relate the drawn figure to an independent measure of body image. Since Buck (1948) and others have suggested that the drawn figure may represent the subject's "body ideal" as well as or instead of his body image, a measure of this construct was also included in the study.

PROCEDURE

The subjects were 30 boys and 30 girls, selected from a total sample of 97 sixth grade public school children so as to equate for age (mean 11-7, range 11-2 to 12-2). The conventional drawing procedure was followed during a regular class session. The children were first asked to draw a person—a whole person, and then a person of the opposite sex

¹ Formerly at the Psychiatric Institute, University of Maryland, where the data for this study were collected.

² As reanalyzed by the present authors, the data presented by Berman and Laffal do *not* show a significant relation between the somatotype of the drawn figures and that of the subjects ($\chi^2 = 6.53$, $df = 4$, $p > .05$).

³ Silverstein and Klee (1958) conducted an experimental test of the body image hypothesis in which body structure was not the basis for selecting subjects, but in this study, too, no independent measure of body image was employed.

from the first. No time limits were set for the task, and no further instructions were given.

When the drawings had been completed, a brief questionnaire was administered. To obtain measures of body image which would be experimentally independent of measures of body structure, the children were asked to estimate their height and weight. For measures of body ideal, they were asked to state how tall they would like to be and how much they would like to weigh (at the present time) if their height and weight could be changed. Finally, the children were weighed and their heights measured.

The height of each of the drawn figures was measured to the nearest 0.1 inch; and an estimate of its volume—the height of the figure multiplied by the square of its width at the waistline (also measured to the nearest 0.1 inch)—was taken to represent its "weight."

RESULTS

The first step in analyzing the data was to assess the degree of correspondence between body image and body structure. For this purpose, Pearson product-moment correlations were calculated between estimated and actual measures, and the significance of the differences between the means of these measures was evaluated using the *t* test for correlated measures. The findings are shown in Table 1.⁴

TABLE 1
CORRELATIONS AND DIFFERENCES BETWEEN
ESTIMATED AND ACTUAL MEASURES
(*N* = 60)

Measure	Estimated		Actual		<i>r</i>	<i>t</i>
	Mean	<i>SD</i>	Mean	<i>SD</i>		
Height in inches	58.3	3.48	59.4	2.92	.86**	4.57**
Weight in pounds	92.0	16.37	94.1	18.56	.95**	2.74*

* Significant at .01 level.

** Significant at .001 level.

The magnitude of the correlation coefficients indicates a rather close correspondence between body image and body structure as far as height and weight are concerned, the correspondence being particularly close in the case of weight. Since the correlations are not perfect, however, there is at least the possibility that estimated and actual measures may be differentially related to measures of

⁴ Data for boys and girls were initially analyzed separately, but results for the two groups were virtually identical, and so they have been combined in the interest of economy of presentation.

TABLE 2
CORRELATIONS BETWEEN ESTIMATED, ACTUAL, AND
IDEAL MEASURES, AND MEASURES OF SAME-SEX
AND OPPOSITE-SEX FIGURES
(*N* = 60)

Figure	Estimated	Actual	Ideal
Height			
Same-Sex	-.33**	-.22	-.09
Opposite-Sex	-.21	-.11	.04
Weight			
Same-Sex	-.27*	-.23	-.19
Opposite-Sex	-.13	-.12	-.18

* Significant at .05 level.

** Significant at .01 level.

the drawn figures (McCornack, 1956). The data of Table 1 reveal a significant tendency for the children to underestimate their height and weight. The most parsimonious interpretation of this finding appears to be that in this period of rapid growth, the children's knowledge of their height and weight is soon outdated.

The correlations between estimated, actual, and ideal measures, on the one hand, and measures of the same-sex and opposite-sex figures, on the other, are given in Table 2. The two coefficients which reach the conventional criteria of statistical significance represent correlations between estimated height and weight, and the estimated height and weight of the same-sex figure. Contrary to expectations, however, both of these coefficients are negative, a finding which at face value suggests an inverse relation between body image and the drawn human figure!

DISCUSSION

It is not at all clear why the subject's mental picture of his physique should be *inversely* related to the physique of his drawn figure. We are reluctant to invoke such concepts as compensation, reaction formation, or contrast projection, for there seems to be no theoretical basis for attempting such "dynamic" interpretations of the present results; nor do we possess additional information on the subjects of this study which might provide independent support for inter-

pretations of this kind. Under these circumstances, we prefer to offer no ad hoc explanation for the findings that estimated measures proved to be somewhat better predictors than did actual measures, and that the relations observed held for same-sex but not for opposite-sex figures. Whatever the interpretations of the data, it is clear that they are not consistent with a body image hypothesis which calls for a direct representation of the body image in human figure drawings.

SUMMARY AND CONCLUSIONS

Human figure drawings were obtained from 60 sixth grade boys and girls, after which they were given a questionnaire designed to provide measures of body image and body ideal. Finally, the children were weighed and measured. Estimated height and weight were highly correlated with actual height and weight, indicating a close correspondence between body image and body structure. When actual, estimated, and ideal measures were correlated with corresponding measures of the drawn figures, small but significant *negative* correlations were obtained between estimated measures and measures of the figures. None of the other

correlations was significant. Without more data than are presently available, these findings are difficult to interpret, but in any case they are not consistent with the assumption that the drawn figure directly represents the subject's body image.

REFERENCES

- BERMAN, S., & LAFFAL, J. Body type and figure drawing. *J. clin. Psychol.*, 1953, 9, 368-370.
- BUCK, J. N. The H-T-P technique: A qualitative and quantitative scoring manual. *J. clin. Psychol.*, 1948, 4, 317-396.
- KOTKOV, B., & GOODMAN, M. The Draw-a-Person tests of obese women. *J. clin. Psychol.*, 1953, 9, 362-364.
- MCCORNACK, R. L. A criticism of studies comparing item-weighting methods. *J. appl. Psychol.*, 1956, 40, 343-344.
- SCHILDER, P. *The image and appearance of the human body*. New York: International Univer. Press, 1950.
- SILVERSTEIN, A. B., & KLEE, G. D. A psychopharmacological test of the "body image" hypothesis. *J. nerv. ment. Dis.*, 1958, 127, 323-329.
- SILVERSTEIN, A. B., & ROBINSON, H. A. The representation of orthopedic disability in children's figure drawings. *J. consult. Psychol.*, 1956, 20, 333-341.
- SWENSEN, C. H., JR. Empirical evaluations of human figure drawings. *Psychol. Bull.*, 1957, 54, 431-466.

(Received March 11, 1960)

TRAINING AND FIRST GRADERS' ACHIEVEMENT

JUNE ELIZABETH CHANCE

University of Missouri

Psychological research regarding parental attitudes and hearing patterns has been mainly concerned with their general effects on personality development of the child. Studies of specific effects of parent-child interaction on the child's academic achievement are rarer, although one might suppose that if variations in parent attitudes and practices produce variation in child personality, these variations would be reflected in—and might in part account for—variations in school achievement.

Results of existing studies of influence of parental variables on school performance suggest, but not unequivocally, that deviations in parent-child relationships are related to deviations in school achievement (Hattwick & Stowell, 1936; Kurtz & Swenson, 1951; Levy, 1933, 1943). These studies, however, are based upon situations where there are extremes in either child or parent behavior and reveal little about what might be true of more typical situations.

Hypotheses offered by McClelland (1958) regarding the role of early learning in formation of the achievement motive suggest relationships between certain facets of parental behavior and the child's motivation toward his school performance. Winterbottom (1958), in a study within the McClelland framework, found that earlier demands by mothers for independence behaviors were related to higher need achievement in 8-year-old boys. She did not, however, find differences between her groups in actual school achievement as assessed by teacher's ratings. Her study, focused as it was on achievement fantasy, did not provide necessary controls of the other pertinent variables which might affect actual achievement.

In a recent study of mothers of deaf

children, Gordon (1959) explored the relation of mothers' independence training attitudes to disparities between the children's intellectual ability and the extent to which they had actually accomplished certain developmental tasks. He found that mothers of high-potential-accomplishment-disparity deaf children favored earlier independence training more than did mothers of low disparity children.

d'Heurle, Mellinger, and Haggard (1959) in a study of personality, intellectual, and achievement patterns of gifted third grade children found small positive correlations between overprotectiveness of parents and arithmetic, reading, and general achievement scores. They also found a positive relationship between parental pressures toward achievement and achievement test scores.

The present investigation asked whether within a group of first grade children, who were not retarded academically nor disturbed emotionally, were individual differences in school progress related to differences in mothers' attitude toward independence training? The study was performed as part of a larger assessment program of first grade children.¹ Medical and psychiatric data, as well as psychological and achievement test results were available for each child. School records, the supervisor of instruction, teachers, and the physician supplied general information about and impressions of home situations. In addition, for the children of the

¹ These data were collected while the author was at the University of North Carolina. The author wishes to express her appreciation to Bernice Wade, Supervisor of Instruction for the Chapel Hill Public Schools, to the teachers, and to Kempton Jones and John Filley whose cooperation made this study possible.

total sample who are included in the present investigation, a questionnaire filled out by the mother regarding her age expectations for achievement of independence behaviors was obtained. The Winterbottom (1958) questionnaire was adapted for this purpose.

METHOD

The children in the first grade class at School A numbered 13 girls and 17 boys; the children at School B, 19 girls and 16 boys. Only children entering first grade for the first time were included in the study. Of the 65 children available for study, 37 were from homes where one or both parents were engaged in occupations related to the local university. Parents of the remaining children were engaged in occupations ranging from professions to owners or managers of small businesses and skilled workers.

All children were administered the Revised Stanford Binet Scale, Form L, between December 1 and April 1 of the school year. In May of that year, both classes were given the Stanford Achievement Test, Primary Battery, Form N. The psychiatrist interviewed each teacher about each child in her class in order to arrive at a psychiatric evaluation. The physician compiled available data to obtain a picture of the medical history and current physical status of each child. Each teacher described all of her children by means of a behavioral checklist. In April, the present investigator mailed to each mother of the children being studied a brief questionnaire composed of some identifying questions, the 20 items of Winterbottom's independence training questionnaire, and 8 other items of a similar type devised by the investigator. A covering letter told mothers they had been selected as a sample to represent mothers of school-age children in the community. Instructions emphasized confidentiality and research use of the data. Mothers were asked to return the questionnaire even if they felt they could not answer every item. They were asked to indicate for each item of the questionnaire the approximate age at which they expected their children to do what the item described and also to check the items they felt were especially important goals of their child rearing.²

The covering letter invited mothers to phone the investigator if they needed further clarification of the questionnaire. Of the five mothers who called, in only one instance did there seem to be genuine confusion about the questionnaire itself. The others seemed merely curious or in need of reassurance. Of 65 questionnaires sent, 52 were returned. The results reported here are based on data from these 52 complete cases.

² Verbatim copies of the instructions and covering letter are available from the author on request.

TABLE 1

MEANS AND STANDARD DEVIATIONS OF INTELLIGENCE, ACHIEVEMENT, AND INDEPENDENCE TRAINING MEASURES FOR ALL SUBJECTS

Measure	Group	N	Mean	SD
Intelligence Quotient (Revised Stanford-Binet, Form L)	Girls	24	124.75	12.89
	Boys	28	130.36	8.77
	Both	52	127.77	11.44
Reading Achievement (Stanford Achievement Test, Primary Battery, Form N)	Girls	24	2.24	.28
	Boys	28	2.55	.55
	Both	52	2.41	.47
Arithmetic Achievement (Stanford Achievement Test, Primary Battery, Form N)	Girls	24	2.18	.34
	Boys	28	2.76	.52
	Both	52	2.58	.49
Revised Independence Training Questionnaire	Girls	24	5.96	1.28
	Boys	28	5.86	1.33
	Both	52	5.90	1.31

TREATMENT OF DATA AND RESULTS

Table 1 presents a summary of means and standard deviations of the measures available for each of the 52 subjects. No child in this group had any marked physical or sensory handicap, nor were any of these children from homes broken by death or divorce. Observation and psychiatric evaluation suggested that 15 children in the total group of 65 showed some mild degree of personality disturbance. Two of these children are included in the 52 of this study—both appear in the early independence training subgroup.

Examination of Table 1 reveals that the group studied was superior in both intellectual ability and school achievement. Tests of mean differences in intelligence scores, reading and arithmetic achievement scores, and independence training scores indicated no significant differences between children in the two different schools, hence the two school groups are combined in Table 1 and in all further analyses of the data.

Since 8 a priori items had been added to the 20 items of Winterbottom's independence training questionnaire, an item analysis of the whole questionnaire was performed. Mothers were divided into groups favorable to early or late training on the basis of their average age of demand scores for the entire questionnaire, i.e., the sum of all ages given divided by the number of items answered. (Not all mothers answered all items; all answered at least 23 of the 28.) Age estimates for each item given

TABLE 2

RANGES AND MEDIANS OF AGE ESTIMATES GIVEN BY ALL MOTHERS TO ITEMS OF THE INDEPENDENCE TRAINING QUESTIONNAIRE AND SIGNIFICANCES OF DIFFERENCES IN RESPONSE BETWEEN MOTHERS FAVORABLE TO EARLY AND THOSE FAVORABLE TO LATE TRAINING

Item	Range of Age Estimates	Median Age Estimate	Value ^a of χ^2
1. To stand up for his own rights with other children	2- 6	3	.36
2. To know his way around his part of the community so he can play where he wants without getting lost	2-10	5	2.84*
3. To go outside to play when he wants to be noisy or boisterous	2- 6	3	.71
4. To be willing to try new things on his own without depending on his mother for help	1- 8	4	21.33****
5. To be active and energetic in climbing, jumping, and sports	1- 8	4	4.06**
6. To show pride in his own ability to do things well	1-10	4	3.98**
7. To take part in his parents' interests and conversations	3-17	7	2.95*
8. To try hard things for himself without asking for help	3-11	6	19.53****
9. To be able to eat alone without help in cutting and handling food	1- 9	6	8.39***
10. To be able to lead other children and to be able to assert himself in children's groups	3-10	5	16.48****
11. To be able to lead other children and to be able to assert himself in children's groups	2- 6	4	17.73****
12. To make his own friends among children of his age	3- 8	6	.31
13. To hang up his own clothes and to look after his own possessions	5-12	6	.76
14. To do well in school on his own	3- 8	6	1.30
15. To be able to undress and to go to bed by himself	2-10	5	9.62***
16. To have interests and hobbies of his own—be able to entertain himself	6-21	10	1.80
17. To earn his own spending money	3- 8	6	2.84*
18. To do some regular tasks around the house	5-16	10	8.39***
19. To be able to stay at home during the day alone	3-15	9	13.20****
20. To make for himself decisions like choosing his clothes or how to spend money for toys, hobbies, recreations, etc.			

TABLE 2—(Continued)

Item	Range of Age Estimates	Median Age Estimate	Value ^a of χ^2
20. To do well in competition with other children—to try hard to come out on top in games and sports	1-10	6	15.75****
21. To be satisfied to stay with someone he knows well when parents must be away for a few days	1- 6	3	.09
22. To decide upon and to purchase small gifts with his own money for family members and close friends	3-10	7	10.48***
23. To hold short conversations with grown-up friends who come to visit the family	4-10	5	4.63**
24. To visit and to stay overnight with a playmate	5- 9	7	.89
25. To straighten out most of his difficulties with other children without adult intervention	4-10	6	3.37*
26. To be interested in obtaining good grades in school	6-12	7	3.96**
27. To take part in group activities such as clubs, scouts, etc.	6-11	8	5.48**
28. To read a simple story or comics by himself	6- 8	7	.14

^a All values of chi square were computed with Yates correction for continuity.

* $p < .10$.

** $p < .05$.

*** $p < .01$.

**** $p < .001$.

by mothers favorable to early training were then classified as above or below the median age estimate given by all mothers for that item. Age estimates given by mothers favorable to late training were similarly classified. Significances of differences in responses to items by the two groups of mothers were tested by means of 2×2 chi squares. The complete list of the independence training questionnaire items is given in Table 2, along with ranges and median age estimates for the total group of mothers. Table 2 also contains the values of chi square and the significance levels of those items answered differently by mothers favorable to early independence training and by those favorable to late. Rescoring each mother's questionnaire using only the responses given to the

19 items showing significant differentiation between the two groups of mothers a revised independence training score (RIT) was obtained. The RIT scores summarized in Table 1 are average age of demand scores, i.e., the sum of estimates given divided by the number out of the 19 items responded to.

In order to investigate the effects of independence training on achievement it was necessary to hold constant differences in intellectual ability. To do this, the distributions of intelligence scores and of reading and arithmetic achievement scores were converted into ranks and thence into standard rank scores. Using these converted scores, a disparity score was computed for each child in reading and in arithmetic. That is, the difference between his ranking in the total

group on the basis of each of his achievement scores and his ranking in the group on the intelligence measure was obtained. If a child's intelligence rank and achievement rank were equal his score would be zero. If his intelligence rank exceeded his achievement rank his score would be negative, or if his achievement rank exceeded his intelligence rank his score would be positive. In order to remove the negative scores a constant of 30 was added to the scores summarized in Table 3. In Table 3, therefore, scores below 30 indicate achievement rank below intelligence rank, while scores above 30 indicate the converse.

Subjects were then divided into early and late independence training groups on the basis of mothers' scores on the RIT questionnaire and mean differences in disparity scores were tested using Fisher's *t*. The analyses were done first for girls and boys separately and then for combined groups. The results of these tests appear in Table 3.

Differences between early and late independence training groups are statistically significant in every case and in a direction suggesting that children whose mothers favor earlier demands for independence make poorer school progress relative to their intelligence level than children whose mothers favor later independence demands. The differences appear, at least superficially, to be more marked in girls than in boys and more marked in reading than arithmetic.

TABLE 3

DISPARITY BETWEEN GROUP RANKING IN INTELLIGENCE SCORES AND RANKING IN ACHIEVEMENT FOR EARLY AND LATE INDEPENDENCE TRAINING GROUPS

Achievement	Early Group		Late Group		Value of <i>t</i>
	Mean	SD	Mean	SD	
Reading					
Girls (24)	20.50	9.98	38.50	9.50	4.32***
Boys (28)	25.43	9.43	35.29	8.06	2.81**
Both (52)	23.15	9.84	36.77	9.07	15.76***
Arithmetic					
Girls (24)	19.33	16.33	37.67	7.89	3.22**
Boys (28)	26.71	10.52	35.14	4.53	2.60*
Both (52)	23.31	14.29	36.31	6.69	4.10***

Tests of the frequency of items checked by mothers on the questionnaire as specially important to them failed to show any significant relation to other measures. Similarly an hypothesis that early and late training groups might contain differential numbers of children from university-related homes was not substantiated.

DISCUSSION

While the findings of this study confirm those of Gordon (1959) with mothers of deaf children, they are contrary to those which might be expected from Winterbottom's study (1958) and from McClelland's (1958) hypotheses about age of independence training and *n* Achievement. However, this study concerns actual achievement, while Winterbottom and McClelland are concerned with motivation to achieve. The multitude of variables which operate and interact to produce differences in actual achievement are exceedingly complex. However, the fact that mothers' attitudes toward earliness of independence training and actual achievement are inversely related in these findings where healthy, psychologically sound youngsters⁸ from a fairly homogeneous social group were studied suggests that relations among maternal attitudes toward independence training, *n* Achievement, and actual achievement need to be more extensively investigated.

If one may generalize from this group of "normal" children to children referred to clinics for academic difficulties, and if one assumes that very early demands for independence behaviors by the mother may be experienced by the child as excessive pressure, the findings here confirm the common clinical hypothesis that school failure is a means by which the child can express resistance to the parent (Vorhaus, 1946).

Examination of the independence training questionnaire from the point of view of what other than attitudes toward independence

⁸ Reanalysis of the data, excluding the two cases of "maladjusted" children from the early training group, failed to alter the results significantly. One, a girl, with an IQ of 146, had a reading disparity score of 19 and an arithmetic disparity score of 18; the other a boy, with an IQ of 125, had a reading disparity score of 36 and an arithmetic disparity score of 28.

* $p < .02$.
 ** $p < .01$.
 *** $p < .001$.

training it might sample suggests that many items could be interpreted as related to the mother's need to maintain interpersonal distance in her relationship with her child. As noted earlier, Kurtz and Swenson (1951) found evidence suggesting that parents of underachievers might be more distant in their relationships with their children than parents of overachievers. d'Heurle, Mellinger, and Haggard (1959) found *both* parental overprotectiveness and pressures for achievement to be positively associated with high achievement. Taking all this evidence together the following hypothesis is suggested: mothers' attitudes toward early independence training will differentially influence the child and his subsequent school achievement depending upon whether she maintains a close or distant interpersonal relationship with him. The same hypothesis also can be stated that maternal attitudes favoring early independence will have different impact upon the child depending upon mother's motivation for desiring that independence. A possible theoretical distinction between instrumental act independence and emotional independence is implied.

The findings of this investigation are limited by the selective nature of the situation studied: the children's superior intellectual ability and small numbers, their attendance in a good school system with excellent teachers, and their parents' relatively high socioeconomic and cultural status. A further source of bias is evident in the differential return of the mail questionnaires. While it is fascinating that of 13 mothers who did not return their questionnaires, all 13 were mothers of children judged by the investigators to be somewhat maladjusted, it is difficult to see how this bias might have influenced the direction of the findings. Rather, it seems to define them all the more clearly.

SUMMARY

The relationship between mothers' attitudes toward independence training and 52 first grade children's school achievement was studied. It was found that children whose mothers favored earlier independence training made less adequate school progress in both reading and arithmetic relative to their intellectual ability than children whose mothers favored later independence training. Implications and limitations of the findings are discussed.

REFERENCES

- D'HEURLE, ADMA, MELLINGER, JEANNE C., & HAGGARD, E. A. Personality, intellectual, and achievement patterns in gifted children. *Psychol. Monogr.*, 1959, 73(13, Whole No. 483).
- GORDON, J. E. Relationships among mothers' n achievement, independence training attitudes, and handicapped children's performance. *J. consult. Psychol.*, 1959, 23, 207-212.
- HATTWICK, B. W., & STOWELL, M. The relation of parental overattentiveness to children's work habits and social adjustments in kindergarten and the first six grades of school. *J. educ. Res.*, 1936, 30, 169-176.
- KURTZ, J. J., & SWENSON, E. J. Factors related to over-achievement and under-achievement in school. *Sch. Rev.*, 1951, 59, 472-480.
- LEVY, D. M. Relations of maternal overprotection to school grades and intelligence tests. *Amer. J. Orthopsychiat.*, 1933, 3, 26-34.
- LEVY, D. M. *Maternal overprotection*. New York: Columbia Univer. Press, 1943.
- MCCLELLAND, D. C. The importance of early learning in the formation of motives. In J. W. Atkinson (Ed.), *Motives in fantasy, action, and society*. Princeton: Van Nostrand, 1958.
- VORHAUS, PAULINE G. Non-reading as an expression of resistance. *Rorschach res. Exch.*, 1946, 10, 60-69.
- WINTERBOTTOM, MARIAN R. The relations of need for achievement to learning experiences in independence and mastery. In J. W. Atkinson (Ed.), *Motives in fantasy, action, and society*. Princeton: Van Nostrand, 1958.

(Received March 16, 1960)

MAGICAL THINKING AND ASSOCIATED PSYCHOLOGICAL REACTIONS TO BLINDNESS¹

STANLEY P. ZARLOCK²

University of Buffalo

With the onset of blindness, an individual becomes aware that he can no longer execute certain sensory-motor responses. A reduction in a repertoire of responses often leads to a state of trauma. Loss of vision hinders ordinary locomotion and creates a special problem of orientation to physical objects. During the initial phase of blindness, even eating a simple meal may become a complex task. Thus, social adaptation and acceptance of a physical disability are largely dependent upon the acquisition of new skills and habits. Among many physically disabled persons, however, blindness becomes an intolerable state of affairs and results in a temporary or permanent regression of ego functions.

The wish for the restoration of vision may cause an individual to rely on an earlier mode of ego development where the boundaries between reality and irreality lack adequate distinction. Consistent with Piaget's (1953) description of the "magico-phe-nomenalistic" state of ego development, a blind person may place unusual confidence in miraculous cures and ascribe magical characteristics to persons and objects that he believes may somehow do away with his blindness. Hence, a regression of ego functions may hinder the learning of new and appropriate social habits.

The present investigation was an attempt to shed some light on the nature and extent of these magical beliefs and to determine the psychological characteristics of personality

which are relevant to the process of social adaptation and acceptance of blindness.

Previous authors have dealt with similar problems. In a theoretical paper, Barker, Wright, Meyerson, and Gonick (1953) described the psychological reactions of blind individuals in terms of Lewin's (1951) model of behavior. These authors stated that blindness creates a new psychological region in which the locations of goals are unknown and behavioral routes are unfamiliar. A new psychological region, therefore, becomes a source of frustration, conflict, and anxiety.

Blindness also creates a state of helplessness and dependency. According to Adorno, Frenkel-Brunswick, Levinson, and Sanford (1950), the "anti-democratic personality" has dependency needs that involve feelings of "doubt, uncertainty, and momentary lack of self-confidence." They illustrated that dependency may be associated with "worry about the future, realization of impending danger, and feeling absolutely lost."

Applying psychoanalytic principles to her study, Burlingham (1941) reported the case histories of two blind children. After close examination, she concluded that the loss of sight seriously interfered with the function of the ego to test reality. These blind children frequently engaged in fantasies in which wishes were fulfilled and certain unpleasant aspects of reality were denied.

Similarly, Deutsch (1940) who studied 28 persons born blind noticed a readiness to give up reality and to escape into fantasy. A large proportion of these subjects believed a cure would come from a supernatural power. Thus, Deutsch concluded:

This expectation that the cure would come from a supernatural power belongs partly to the realm of fantasy in which there is room for the fulfillment of all wishes (p. 124).

¹ This paper is a condensation of a dissertation submitted in partial fulfillment of the requirements for the PhD degree at the University of Buffalo. The author is especially indebted to B. Richard Bugelski for his assistance on the project.

² Now a staff psychologist at the Veterans Administration Hospital, Lexington, Kentucky.

For the present study, a series of hypotheses dealing with the relation between the psychological characteristics of the blind and social adaptation were formulated. Among them were the following:

1. Social adaptation to a new psychological region is largely enhanced by ego strength, low manifest anxiety, and a positive attitude towards blindness.
2. Failure at social adaptation results in magical beliefs as to the power of medicine and religion in the treatment of illness and disability.
3. An antidemocratic personality is associated with failure in social adaptation.
4. There are no personality differences between socially well-adjusted blind subjects and physically normal individuals.

METHOD

Subjects

Fifty-two blind subjects from various regions of New York State were studied. Each subject was a male between 20 and 45 years of age. Four additional criteria were used in the selection of subjects: total and permanent blindness with no light perception; at least 10 years of normal vision prior to blindness; blind for at least 3 years; with the exception of blindness, no other physical disability.

Twenty-five physically normal subjects were selected as a control group for the 25 blind subjects who had made an adequate social adjustment to their handicap. These groups of subjects were matched on the variables of age, intelligence, socioeconomic background, and religious affiliation.

Scales

Based on the research of Barker (1948), Bauman (1954), Fitting (1953), Meyerson (1953), and Raskin (1953), a Social Adjustment Scale was developed as a preliminary step to permit measurement of social adaptability to blindness. The scale was a rating device based on a series of check list items which indicated the blind individual's level of social maturity or acquisition of appropriate skills in 10 problem areas: employment, travel, indoor orientation, socialization, communication, recreation, eating problems, dressing problems, business problems, and physical hygiene.

In the development of the Social Adjustment Scale, care was taken that each problem area was a measure of social and not a measure of psychological behavior. For instance, some of the items were as follows: gainfully employed for one year, uses cane or dog guide to travel about the community, reads and writes braille, holds membership

in a civic organization, maintains proper etiquette during meals, and buys his own clothes.

Each blind subject was rated on the Social Adjustment Scale by a social worker most familiar with him. A second rating was obtained by the investigator who interviewed each subject and members of the family. A reliability coefficient of correlation between the two sets of scores was .92. Applying the Spearman-Brown Prophecy Formula, the correlation was raised to .95.

Personal adaptability, manifest anxiety, and attitudes towards blindness were measured by the Barron Ego Strength scale (1953), Taylor Manifest Anxiety Scale (1953), and the Fitting Attitudes towards Blindness Scale (1953), respectively.

Two separate attitude scales were constructed to get at the confidence placed in medicine or religion to restore sight. Each scale consisted of statements which reflected attitudes from extremely negative to extremely positive. On the Religious Scale, for instance, the statement, "The healing of lepers as mentioned in the Bible is a fairy tale," was rated negative towards religion; whereas, the statement, "A perfect spiritual faith would absolutely lift us from all physical disease," was considered extremely positive. For each of the two scales, 22 statements were selected by Thurstone's (1951) method from original lists of 130; each was rated on an 11-point scale.

Both instruments were administered to 80 undergraduate college students. The split-half reliability coefficient of correlation was .80 for the Medical Scale, .93 for the Religious Scale. The Spearman-Brown Prophecy Formula raised the coefficients to .92 and .96.

The California F Scale, Form 78, devised as a test of dependency and reliance on authority was used to measure "antidemocratic personality."

An estimate of intelligence was obtained from the Vocabulary subtest of the Wechsler-Bellevue scale, Form 1 (1944).

At the end of the psychological test battery, an interview supplemented the data obtained from the Medical and Religious Scales. During the interview, an additional attempt was made to determine the attitudes of blind subjects towards "miraculous" cures by allowing the subjects to verbalize their problems.

Procedure

The social worker most familiar with the subject was asked to rate him on the scale of social adjustment to blindness. The ratings of each social worker were corroborated by similar ratings obtained from a member of the family and a close acquaintance. Because the scale dealt with behavioral items, there was little disagreement. Scores could range from 0 to 30, with higher scores indicating superior adjustment.

The sequence of test presentation was always constant. Test items were read aloud to each blind subject and the answers tape recorded. The interview was recorded verbatim and then rated by two

sophisticated judges along a five-point scale of confidence the subject ascribed to medicine and religion for bringing about a miraculous cure. A rating of 1 indicated a strong confidence and a rating of 5, no confidence. The same procedure was followed with physically normal subjects except that they were not rated for social adaptability.

RESULTS

On the Social Adjustment Scale, the scores for the 52 blind subjects ranged from a low of 3 to a high of 29. The mean score for the entire sample was 16.3. The distribution of scores approximated a normal probability curve.

As Table 1 indicates, all psychological measures correlated significantly (.01 level) with scores on the Social Adjustment Scale. These correlations support the hypotheses that higher ego strength, lower manifest anxiety, and a more positive attitude towards blindness are important psychological variables which are strongly related to a blind individual's ability to make an adequate social adjustment. Reliance on the power of medicine and religion to restore vision was negatively correlated with the blind person's level of social adaptation. Antidemocratic personality was also negatively related to social adjustment to blindness. When the possible effects of intelligence and duration of blindness were partialled out, all correlations remained significant at the .01 level with the exception of the one associated

TABLE 1

CORRELATIONS BETWEEN SCORES ON SOCIAL ADJUSTMENT SCALE AND PSYCHOMETRIC TESTS

Score	r^*
Ego Strength Scale	.71
Manifest Anxiety Scale	-.64
Attitude towards Blindness	.53
Attitude towards Medicine	-.61
Attitude towards Religion	-.60
California F Scale	-.62

* All significant at .01 level.

with the Medical Scale for which the confidence level dropped to .03.

A comparison of scores on each of the psychological tests indicated that no significant differences existed between the 25 physically normal individuals and the 25 blind subjects who scored above average on the Social Adjustment Scale.

As can be seen in Table 2, blind subjects who scored below the mean on the Social Adjustment Scale were frequently rated from their interviews as individuals who rejected their blindness and believed in miraculous cures. Better adjusted blind subjects tended to accept their blindness and rejected the idea of a miraculous cure. Physically normal subjects tended to be neutral.

The distribution of scores for both blind groups as indicated in Table 2 shows the

TABLE 2

RATINGS OF THE ATTITUDES TOWARDS MEDICINE AND RELIGION BASED ON INTERVIEW PROTOCOLS

Group	N	Medicine					Religion				
		1	2	3	4	5	1	2	3	4	5
Poorly Adjusted Blind	27	32%	32%	0%	18%	18%	63%	10%	3%	6%	18%
Well Adjusted Blind	25	7%	7%	0%	11%	75%	3%	7%	0%	4%	86%
Visually Normal Subjects	25	20%	20%	36%	12%	12%	24%	28%	28%	8%	12%

Note.—Scale values run from 1 (Positive Towards) to 5 (Negative Towards). Values shown are percentages of ratings by two judges.

divergence of observed from expected results was significant at the .01 level. For the physically normal subjects, the divergence of observed from expected results was not significant at the .05 level.

To test for similarity among psychological tests, a series of intercorrelations were obtained. These intercorrelations ranged from .04 to .59, with the mean correlation at .30. The intercorrelations which fell above the mean were: F Scale and Medical Scale .59; F Scale and Religious Scale .57; and Medical Scale with Religious Scale .38.

DISCUSSION

An examination of the results clearly shows that social adjustment to blindness is closely related to ego strength, manifest anxiety, and attitudes towards blindness. To a large extent, these psychological variables appear to determine the kind of personal and social adjustment an individual makes to his physical disability. On the other hand, the psychological characteristics of an individual may be a reflection of the amount of social adjustment he has made towards his handicap.

The role of ego strength in the process of adjustment to blindness becomes more meaningful when examined in the light of somatopsychology (Barker et al., 1953). Orientation to a new psychological region often requires a change in physical and cognitive behavior. Likewise, adjustment to blindness generally necessitates a new focus on goals and the learning of new skills and habits by which these goals are attained. Thus, it seems reasonable to assume that personal adaptability and resourcefulness have an important bearing on the reaction of an individual to a new psychological region.

The uncertainty of goals and behavioral routes of a new psychological region arouse manifest anxiety. On the other hand, the development of new skills and habits tends to abolish the unfamiliarity of a new region and thus results in less manifest anxiety.

When a person perceives his world as competitive and hostile, lack of vision may arouse in him a negative attitude towards blindness. He may feel that a lack of sight places him at a serious disadvantage with

other human beings. The person may attempt to withdraw from the "struggle" and avoid competition against unfavorable odds.

The individual who perceives the world in a less threatening sense generally has a more favorable attitude towards blindness. The physical disability does not place him in a dangerous position; hence, there is no need to withdraw from social situations. He acquires new skills and habits which permit him to carry out many of his personal and social functions.

As mentioned at the beginning of the study, a blind individual may seek an escape route by which he avoids the unpleasant aspects of the new psychological region. A common escape route, for many of these individuals, takes the form of magical beliefs towards the recovery of vision. The results demonstrate that subjects who resisted the acceptance of blindness placed unusual confidence in the ability of medicine and religion to perform miraculous cures. Subjects who accepted their physical disability expressed more realistic attitudes towards medicine and religion.

Finally, attitudes towards authority have an important bearing on the adjustment process to blindness. Making an inference from the work of Adorno et al. (1950), it is probable that the blind individual who has an antidemocratic personality finds the new psychological environment full of dangerous elements. The overwhelming threat may cause him to lose self-confidence and depend heavily on the support of authority as a means of survival. The less authoritarian-minded person probably does not perceive the environment quite as threatening. Since extremely dangerous elements are not present, he is able to maintain self-confidence and achieve levels of independent behavior.

In the area of magical thinking, intelligence had some influence on the formation of attitudes towards miraculous cures in medicine. However, intelligence had almost no effect on attitudes towards miraculous cures in religion.

Many intelligent blind individuals eventually adopt realistic attitudes towards medicine. However, if acceptance of blindness does not accompany a realistic attitude

towards medicine, these individuals may discover new areas into which they project their magical beliefs. Having failed to regain vision, they may shift their magical beliefs from ophthalmology to religion or to some other frame of reference through which they hope to experience a miraculous cure.

A comparison of the F Scale with both the Medical and Religious Scales revealed comparatively high correlations. These correlations suggest the presence of a common psychological factor. The three scales undoubtedly measure personal reactions to authority. Medicine and religion could be considered as specific domains of authority; whereas, items on the F Scale pertain to authority in a general sense. Thus, attitudes towards authority in the broad sense may be identical with attitudes towards authority in specific areas.

SUMMARY

Fifty-two blind subjects were rated on the Social Adjustment Scale and then tested for ego strength, manifest anxiety, attitudes towards blindness, attitudes towards the efficacy of medicine and religion to restore health, and degrees of antidemocratic personality. Each blind subject was interviewed so that additional data could be obtained about attitudes towards a miraculous cure of blindness. Twenty-five physically normal subjects were matched with 25 blind subjects who scored highest on the Social Adjustment Scale. The control subjects followed the same procedure as that of the blind subjects but were not rated for social adaptability.

The results indicated that social adaptation to blindness was related to high ego strength, low manifest anxiety, and a positive attitude towards blindness. Blind subjects who were socially maladjusted placed unusual confidence in the efficacy of medicine and religion to restore health and strongly

believed they would regain vision through a miraculous cure. Subjects who were poorly adjusted to the environment were characterized as more antidemocratic in their personalities than blind subjects who had made a good social adjustment to their handicap.

No significant differences were obtained between socially adapted blind and physically normal subjects.

REFERENCES

- ADORNO, T. W., FRENKEL-BRUNSWICK, ELSE, LEVINSON, D. J., & SANFORD, R. N. *The authoritarian personality*. New York: Harper, 1950.
- BARKER, R. G. The social psychology of physical disability. *J. soc. Issues*, 1948, 4(4), 28-38.
- BARKER, R. G., WRIGHT, B. A., MEYERSON, L., & GONICK, M. R. *Adjustment to physical handicap and illness: A survey of the social psychology of physique and disability*. (Rev. ed.) New York: Social Science Research Council, 1953.
- BARRON, F. An ego-strength scale which predicts response to psychotherapy. *J. consult. Psychol.*, 1953, 17, 327-333.
- BAUMAN, M. K. *Adjustment to blindness*. Harrisburg, Pa.: State Council for the Blind, 1954.
- BURLINGHAM, D. Psychic problems of the blind. *Amer. Imago*, 1941, 2, 43-85.
- DEUTSCH, F. The sense of reality in persons born blind. *J. Psychol.*, 1940, 10, 121-140.
- FITTING, E. A. *Evaluation of adjustment to blindness*. New York: American Foundation for the Blind, 1953.
- LEWIN, K. *Field theory in social science*. New York: Harper, 1951.
- MEYERSON, L. The visually handicapped. *Rev. educ. Res.*, 1953, 23, 476-491.
- PIAGET, J. *The construction of reality in the child*. New York: Basic Books, 1953.
- RASKIN, N. J., & WELLER, M. F. *Current research in work for the blind*. New York: American Foundation for the Blind, 1953.
- TAYLOR, J. A. A personality scale of manifest anxiety. *J. abnorm. soc. Psychol.*, 1953, 48, 285-290.
- THEURSTONE, L. L., & CHAVE, E. J. *The measurement of attitude*. (2nd ed.) Chicago: Univer. Chicago Press, 1951.
- WECHSLER, D. *The measurement of adult intelligence*. Baltimore: Williams & Wilkins, 1944.

(Received March 17, 1960)

SPEECH DISTURBANCE AND JUDGED ANXIETY

DONALD S. BOOMER AND D. WELLS GOODRICH¹

National Institute of Mental Health

In the comparatively brief history of systematic research in psychotherapy, verbal measures have been given a great deal of attention. Much of this work has taken the form of developing systems of categories for classifying manifest content. A survey of this literature by Auld and Murray (1955) lists 99 content analysis studies, all of them dealing in one way or another with *what* people say.

Although it is generally recognized that *how* people talk is also a significant aspect of their speaking behavior, comparatively little attention has been devoted to the formal, structural, and expressive aspects of speech. The major variables which have been systematically studied include: rate of speech (Goldman-Eisler, 1956; Henze, 1953; Lasswell, 1935), the interaction chronograph variables developed by Chapple (1939), microlinguistics (McQuown, 1957; Pittenger & Smith, 1957), and speech disturbances (Dibner, 1956; Mahl, 1956).

The present paper is based on a replication of the Mahl study cited above. Since this study will only be summarized here, interested readers should consult the original report for details. In brief, Mahl presented an empirically derived set of categories of disturbance which occur frequently in speech. The categories are: (a) "ah," (b) sentence correction, (c) sentence incompleteness, (d) repetition, (e) stutter, (f) incoherent sound, (g) tongue-slip, (h) omission of a word or part of a word. In a subsequent report Mahl (1959) excluded "ah" from this list. Our own experience leads us to concur in this

exclusion; hence the speech disturbance data in this paper will refer only to categories *b-h*, which Mahl terms the "non-ah" disturbances.

These speech disturbances (hereinafter referred to as SDs) are scored by listening to recordings of speech and marking each disturbance on a typewritten transcript at the precise place in the text where it occurs. The speech disturbance ratio (SDR) for any selected passage can then be calculated by dividing the number of speech disturbances by the number of words spoken.

Mahl hypothesized that SDR varies directly with fluctuations in the speaker's anxiety level. As a test of this hypothesis he compared SDRs with his clinical judgments of anxiety in six therapeutic interviews with a patient he had treated 2½ years previously. Contamination was avoided by having all SDs edited out of the typewritten transcripts from which the clinical judgments were made. His central finding was that the mean SDR was significantly higher for those phases of the interviews he judged "high anxious" than for those phases he judged "low anxious."

At the time Mahl's study was published the present authors were trying to define and measure some aspects of patients' speech which might reflect the immediate effects of specific therapeutic interventions. The SDR appeared to be a potentially useful research instrument for this purpose, provided that certain issues could be clarified. These issues are embodied in the following hypotheses which were the basis of this replication and extension of the Mahl study:

H₁. High anxiety interview phases, as judged by the therapist, have a greater mean SDR than low anxiety interview phases.

¹We gratefully acknowledge the contributions made to this research by several colleagues: Janet Barclay, principal research assistant; Allen T. Dittmann, Leonard Duhl, and Joseph Handlon, clinical judges; Sam Greenhouse, statistical consultant; and Charles Odell, research assistant.

This, of course, was Mahl's fundamental finding. Our aim was straightforward replication: could this effect be successfully demonstrated again using different patients and therapist-judges?

H₂. High anxiety interview phases, as judged by persons other than the therapist, have a greater mean SDR than low anxiety interview phases.

This hypothesis formalizes the question as to whether the judgment of anxiety requires the intimate, detailed, firsthand knowledge of the patient possessed only by the therapist. This question cannot be answered from Mahl's results since he was both therapist and sole judge.

H₃. Independent judgments by different judges demonstrate substantial agreement as to the identification of phases of high and low anxiety.

This is a test of the interjudge reliability of the criterion judgment of anxiety. It also tests one of the links in the logical chain which relates the SDR to anxiety, since the statement that A is a measure of B requires that values for both A and B can be determined with acceptable reliability and that they covary systematically. Mahl's study demonstrates only that the SDR is a reliable measure and that it covaries with his assessment of anxiety. The present study has sought evidence concerning the missing term, that is, the reliability with which fluctuations of anxiety within an interview can be judged.

METHOD

Subjects. The subjects for this study were two outpatients in a psychotherapy research project. Mrs. Alpha had been seen for approximately 150 hours by one of the authors. Mrs. Beta had been seen for approximately 50 hours by the other author. All of their interviews had been tape recorded.

Procedure. For each patient four interviews were selected randomly and transcribed verbatim for SD scoring. Two research assistants were trained to score the SDs and did all of the scoring for this study. The reliability of their scoring approximated the .94 figure reported by Mahl. After scoring, the transcripts were retyped with all SDs edited out. The judgments of anxiety were made from these edited transcripts.

Anxiety Judgments. These judgments were made according to the procedure described by Mahl in his report and in personal discussions with the authors. The preparation for the judging task began with an extended clinical review of the case. The judges then listened to and discussed a selection of the recorded interviews, including those interviews immediately preceding and following the test interviews, but excluding, of course, the test interviews themselves.

After this preparation, which required about 15 hours, the judges independently made the clinical judgments of anxiety on the test interviews. The task was to divide each interview into a series of phases, representing periods of relatively high or relatively low anxiety for the patient. In describing this judging task Mahl (1956) says:

During therapeutic sessions and while studying recordings it often appears that interviews are divisible into "natural" segments or phases, each of which could be assigned to a single theme of content or interaction, and that the patient becomes anxious and conflictful in some, but becomes less anxious in others (p. 6).

The judging procedure, as described by Mahl, requires what can best be termed "immersion" in the interview material. The judges read and reread the typescripts at odd moments over a period of several weeks, making notes, revising, and trying out tentative approaches until the emerging phases seemed to stabilize. Finally these phases were marked off in the transcript and labeled "high" or "low" anxiety. The clinical preparation and judging were carried through to completion for Mrs. Alpha before beginning with Mrs. Beta.

The judges were the authors, who judged both cases, and three additional volunteers, two of whom judged Mrs. Beta's interviews and one of whom judged Mrs. Alpha's. All five judges were trained psychotherapists with 5 to 10 years' experience.

RESULTS

The data to be presented in this section will be organized in terms of the three hypotheses listed in the introduction.

H₁. Only the phase judgments of the authors on their respective patients were relevant to this hypothesis. The data analysis was directly adopted from Mahl. For each phase of each interview a SDR was computed, and mean SDRs calculated for the high anxiety and low anxiety phases for each patient.²

² For individual interviews the number of high and low phases were not always equal. In order to avoid the bias of the "hour effect" described by Mahl (1956, p. 7) the excess number of high or low phases were randomly discarded from those interviews in which the numbers were not equal.

The results of this test are equivocal. In the case of Mrs. Beta the high anxiety phases established by her therapist had a significantly higher mean SDR than did the low anxiety phases ($t = 2.29$; $p < .05$). In the case of Mrs. Alpha there was no significant difference between these means.

H_2 . In the case of Mrs. Alpha, it will be recalled, two judges besides her therapist judged the interviews; for Mrs. Beta there were three additional judges. The data analysis was the same as that for H_1 . This hypothesis was not supported. None of the five sets of judgments by the nontherapists showed the predicted SDR discrepancy.

H_3 . It must be stated at the outset that this hypothesis could not be directly tested without altering the design such that it would no longer constitute a replication. The inherent statistical obstacle is that the free judgment instructions require the judges to establish their own units, i.e., motivational phases, which they are to label high or low anxiety. Since judges can and do differ with one another regarding the number and limits of the phases they discern in a given interview, there are no common units on which to base a quantitative statement about their agreement.

Furthermore, no arbitrary division of the interview into comparison units, minutes, say, would serve the purpose. Such minute-units could not defensibly be regarded as independent events for statistical analysis because the original judgments had been made on groups of from 2 to 20 consecutive minutes.

Since a direct test was impossible, a method was devised to permit a statement by indirection about interjudge agreement. The reasoning was as follows: (a) In attempting to specify interview phases of high and low anxiety, judges agree on some passages and disagree on others. (b) If the passages of agreement are not random, but do, in fact, reflect the judges' consensus about the level of anxiety, Mahl's hypothesis would predict systematic SDR differences between consensually high and low anxiety passages. (c) If the high and low consensual passages are not different with regard to SDR one of two things is true: interjudge agreement,

where it exists, does not represent agreement-about-anxiety, or else the SDR does not measure anxiety.

The statistical procedure was as follows: Transcripts of the eight test interviews were divided into minutes. These minute-units were then consecutively tabulated in terms of their labeling as high or low anxiety by all judges. Consensual "phases" were constructed by selecting all sequences of 2 or more consecutive minutes in which all judges, or all judges but one agreed. It was possible thus to construct eight high and eight low consensual phases for each patient. These constructed phases were comparable in length to the phases which had emerged from the individual judgments since they averaged 6 minutes with a range of from 2 to 19 minutes. In the aggregate, these phases were a large sample, representing more than 50% of the interview material.

Analysis revealed that for neither patient did the high anxiety and low anxiety consensual phases differ with regard to mean SDR. In terms of the argument presented above, these data furnish no support for the hypothesis that consensus among judges represents consensus about anxiety, if anxiety is presumed to be accompanied by high SDRs. Further elaboration of this point will be reserved for the discussion section.

DISCUSSION

The results of this study can be summarized as follows: for one of the two patients Mahl's finding was successfully replicated, i.e., her therapist's judgments of periods of high and low anxiety were positively and significantly related to the SDR during these periods. Anxiety judgments on the same interview material made by three additional judges failed, however, to show any significant relationship to the SDR.

For a second patient her therapist's anxiety judgments and those of two additional judges uniformly failed to show any relationship to SDR. Finally, those sections of both patients' interviews identified by the majority of the judges as "more anxious" showed no higher SDR than those sections consensually labeled "less anxious."

These findings do not lend themselves to

any clear-cut conclusions about the SDR. The following issues, however, have been clarified:

1. The Mahl hypothesis was supported in one of the cases, suggesting that his results have a degree of generalizability which warrants further research.

2. In the other case the findings were negative. This indicates that the SDR cannot be uncritically accepted as a universal measure of intercurrent anxiety in psychotherapy interviews. Further investigation is necessary in order to specify the conditions under which the SDR can be meaningfully employed.

3. The hypothesized relationship between SDR and anxiety judged from transcripts holds, when it does hold, only for the judgments made by the therapist who treated the patient. This was true of Mahl's patient and of Mrs. Beta in the present study. Thus Mahl's criterion, the therapists' judgment, remains essentially private and unrepeatable.

The unfortunate consequences of this state of affairs can be seen in the present study. In seeking to validate a new measure like SDR against a criterion of unknown reliability only positive results are informative. Negative findings leave the issues confused. The failure of replication with Mrs. Alpha, for example, may be interpreted in several equally likely ways. It is possible that her anxiety fluctuations are not reflected in her SDR, or that her anxiety fluctuations are not reflected in her written transcript, or that these particular judges are not sufficiently sensitive to discern her anxiety fluctuations.

These, of course, are not all of the possible interpretations, but they suffice to make the familiar point that one of the ends of a measurement hypothesis requires empirical anchoring. Mrs. Alpha's data, being negative, show that the SDR hypothesis needs refinement; being equivocal, they provide no crucial evidence for any specific refinement.

Further Research

The foregoing discussion indicated the need for more specification regarding the properties and the limitations of the SDR. Also illustrated were the shortcomings of the

research use of clinical judgment of anxiety fluctuations in providing the necessary crucial evidence. No simple solution can be suggested for this problem. No highly reliable measure of anxiety can be offered as a substitute for clinical judgment, since no such measure exists. At our present stage of knowledge the phenomena associated with anxiety can be accounted for only by a complex description which involves observations and inferences about physiological activity, private experience, and observable behavior, both verbal and nonverbal. The complex interrelationships among this loose network of phenomena are not now sufficiently explicit to warrant the belief that anxiety can be scaled along any single dimension or gauged at any one level of functioning.³

In the face of such theoretical complexity the preferred research strategy might be to regard the SDR as a promising measure of certain aspects of anxiety in certain classes of people under certain conditions. Experimental work, foregoing the attempt to demonstrate that the SDR measures anxiety, could focus on some limited prior questions concerning the psychological properties of the measure and the manner and conditions of its covariance with other reproducible and reliable measures which may also be presumed to reflect some aspects of anxiety.

Some research along these lines has already begun. Panek and Martin (1959) building on Mahl's work, have demonstrated with a group of psychotherapy patients that GSR dips are preceded by rising SDR gradients and followed by declining gradients. Dittmann⁴ is studying some temporal relationships between SDR and certain body movements. The present authors are investigating possible relationships between SDR and rate of speech. The strategic advantage of this part-problem approach to anxiety measurement is that the use of reproducible

³ Dibner (1958) intercorrelated five measures of anxiety: skin conductance, patients' self-ratings, clinicians' ratings, and two separate measures of speech disturbance. Of the 10 correlations thus generated, 8 were not significantly different from zero.

⁴ Personal communication, 1959.

methods and measures makes it possible ultimately to integrate these and other similar studies experimentally and conceptually.

SUMMARY

This research was an attempt to repeat a pioneering study by Mahl (1956), in which he demonstrated that the incidence of certain disturbances of speech increased during portions of psychotherapy interviews judged to be anxious, and decreased during portions judged less anxious.

The results of the present test were inconclusive. The anxiety judgments for one patient made by her therapist supported Mahl's finding, but in a second case the judgments made by the therapist failed of replication. The judgments made by five additional judges who were not the patients' therapists uniformly failed to show the hypothesized relationship to the speech disturbance measures. Reconciliation of these results must await further research.

REFERENCES

- AULD, F., & MURRAY, E. J. Content-analysis studies of psychotherapy. *Psychol. Bull.*, 1955, 52, 377-395.
- CHAPPLE, E. D. Quantitative analysis of the inter-

- action of individuals. *Proc. Nat. Acad. Sci., Wash.*, 1939, 25, 58-67.
- DIBNER, A. S. Cue counting: A measure of anxiety in interviews. *J. consult. Psychol.*, 1956, 20, 475-478.
- DIBNER, A. S. Ambiguity and anxiety. *J. abnorm. soc. Psychol.*, 1958, 56, 165-174.
- GOLDMAN-EISLER, FRIEDA. The determinants of the rate of speech output and their mutual relations. *J. psychosom. Res.*, 1956, 1, 137-143.
- HENZE, R. Experimentelle Untersuchungen zur Phänomenologie der Sprachgeschwindigkeit. *Z. exp. angew. Psychol.*, 1953, 1, 214-243.
- LASSWELL, H. D. Verbal references and physiological changes during the psychiatric interview: A preliminary communication. *Psychoanal. Rev.*, 1935, 22, 1-24.
- MCQUOWN, N. A. Linguistic transcription and specification of psychiatric interview material. *Psychiatry*, 1957, 20, 79-86.
- MAHL, G. F. Disturbances and silences in the patient's speech in psychotherapy. *J. abnorm. soc. Psychol.*, 1956, 53, 1-15.
- MAHL, G. F. Measuring the patient's anxiety during interviews from "expressive" aspects of his speech. *Trans. NY Acad. Sci., Ser. 2*, 1959, 21, 249-257.
- PANEK, D. M., & MARTIN, B. The relationship between GSR and speech disturbances in psychotherapy. *J. abnorm. soc. Psychol.*, 1959, 58, 402-405.
- PITTINGER, R. E., & SMITH, H. L. A basis for some contributions of linguistics to psychiatry. *Psychiatry*, 1957, 20, 61-68.

(Received March 24, 1960)

NORMAL, HYPNOTICALLY INDUCED, AND FEIGNED ANXIETY AS REFLECTED IN AND DETECTED BY THE MMPI¹

ALBERT A. BRANCA AND EDWARD E. PODOLNICK²
University of Delaware

The use of hypnosis as a technique for the production of signs of disorder in normal people has been attempted. Luria (1932), and Huston, Shakow, and Erickson (1934) showed that word association techniques together with certain motor responses were successful in revealing the presence of emotion arousing conflicts that had been suggested in hypnosis. Fisher and Marrow (1934) reported significant differences in reaction times obtained in hypnotically induced "moods" of elation and depression. Sweetland (1948) suggested certain psychiatric syndromes to normal subjects who had been hypnotized. Comparison of MMPI profiles obtained when these syndromes were suggested indicated that it was possible to produce "laboratory neuroses" by hypnosis. Grosz and Levitt (1959) suggested anxiety to 12 hypnotized medical and nursing students. They reported increased scores on the Taylor Manifest Anxiety Scale and diminished scores on the Barron Ego Strength scale. They also reported that scores on the two tests taken during the waking state did not differ from scores obtained during hypnotic states when anxiety was not suggested.

Studies also show that the MMPI validity scales can identify dissemblers. Gough (1947) showed that the MMPI was able to identify "fakers" even when they were psychiatrists, clinical psychologists, and social workers who were familiar with the diagnostic signs of behavior disorders as well as the MMPI. Other investigators (Cofer, Chance, & Judson, 1949; Hunt, 1948) also indicate that the MMPI, through separate

or combined use of its validity scores, is capable of differentiating between dissemblers and other groups.

In 1952 Welsh added an Anxiety (A) scale to the MMPI. This development has made it possible to observe the effects of suggesting this simpler and more general symptom of disorder.

The specific hypotheses of this experiment are:

1. There will be a significant increase in the A scale of the MMPI between scores obtained under normal conditions and under conditions of hypnotically induced anxiety.

2. The validity scales of the MMPI will differentiate between profiles obtained under conditions of dissembling and profiles obtained under conditions of both normal and hypnotically induced anxiety.

METHOD

Subjects

Ten students, of whom eight were female, were used as subjects in this experiment. The normal records were obtained from students who had taken the MMPI as part of a classroom demonstration. At the time the first profiles were obtained, the students were not aware that they might be called upon to participate in an experiment. Students from other freshman and sophomore courses volunteered to participate when they had heard about the study. Two of these students were used as experimental subjects. These two were given to believe that the MMPI was being used as a screening device and not a part of the experiment proper.

Experimental candidates were selected on the basis of: (a) normal MMPI profiles and anxiety scores; (b) absence of a history of treatment for mental disorder; (c) absence of a history of epilepsy, or convulsions, or neurological disease of any type; (d) a willingness to participate in the experiment. Actual subjects were selected from this larger group on the basis of hypnotizability. The criterion for

¹ This research was supported by a University of Delaware Faculty Summer Research Grant.

² Now at Bucknell University.

depth of trance was the elicitation of positive auditory and visual hallucinations. Out of a total of 50 experimental candidates, 10 met this criterion, 2 males and 8 females. This percentage is consistent with others also reporting approximately 20% success in obtaining a deep trance (Dorcus, 1956). A disproportionately large number of females volunteered to participate in the experiment.

Procedure

Each hypnotic session was held in a room with a one-way observation screen and an intercom system. In this way one experimenter was able to observe each session while the other performed the hypnosis. Each candidate was made aware of this observation.

The first phase of the experiment consisted of training sessions wherein the experimental candidates were trained to achieve the trance state. When a depth of trance was reached in which positive auditory and visual hallucinations were produced, the candidate met the criterion for inclusion as an experimental subject and an anxiety state was suggested. The instructions for producing anxiety were obtained from definitions and descriptions of anxiety by various authors (Conklin, 1936; Heyns, 1958; Lehner & Kube, 1955; May, 1950; Shaffer & Shoben, 1956; Warren, 1934). These instructions were as follows:

You are beginning to feel very uneasy and anxious. You don't know why, but this uneasy feeling is making you nervous, irritable, and frightened. You feel as if something dreadful is about to happen but you don't know what. This feeling of dread is mingled with a curious feeling of hope that is very unpleasant. You are becoming more and more apprehensive. You are in a state of anxious expectation and self-doubt. You feel now as if you are threatened and it frightens you. You feel as if you are about to lose something important to you, or be hurt. This anxiety is becoming stronger and stronger. Now you feel as if something is wrong, as though you had neglected to do something very important, but you can't recall what it is. You feel, though, that whatever it is, it is making you feel on edge and uneasy. It is making you feel blue, melancholy, unhappy, and excited in an unpleasant way. You feel frightened, but you don't know what it is you are frightened about. This is certainly an unpleasant form of excitement. You are now very apprehensive and anxious.

After the anxiety instructions were read, the MMPI was readministered. The subjects took between 70 and 90 minutes to complete the MMPI. In order to maintain the trance state for that period instructions and suggestions reinforcing the trance state were given when the subject had reached the halfway point in the test. At that time the anxiety instructions were also reread to each subject. The subjects were aroused from the trance state after suggestions counteracting the anxiety were made. These instructions, given twice, were as follows:

You are beginning to feel less apprehensive and anxious. The unpleasant form of excitement caused by the fact that you were frightened is leaving you. You are beginning to feel happier, more alert, and relaxed. You no longer feel on edge or uneasy, and you are experiencing a feeling of well-being. You are now confident and at ease. You feel happy and at peace with the world. You are experiencing a soothing calmness and you feel warm, relaxed, comfortable, and alert. You don't feel nervous, irritable, or frightened any more. You are no longer apprehensive and no longer feel self-doubt. You don't feel as if you are frightened or are about to be hurt. You don't feel as if you're about to lose something but don't know what. You are now very relaxed. You feel as if all of your troubles and problems are leaving you. You feel as if all of your fears are gone and this gives you a feeling of ease and comfort. You are happy and relaxed and normal.

At the hypnotic session, subjects were instructed to remember all events that occurred during the trance state.

In the final phase of the experiment, which took place approximately a week after the first phase, each subject was told that he was to make believe that he was anxious and that he was to "fake" anxiety while taking the MMPI. The same description of the anxiety state was read to him again with the statements "make believe that" or "pretend that" prefixing each sentence. He was further instructed to mark the test as though he were trying

TABLE 1
MMPI T SCORE MEANS AND STANDARD DEVIATIONS
UNDER NORMAL CONDITIONS, UNDER CONDITIONS OF
HYPNOTICALLY INDUCED ANXIETY, AND UNDER
CONDITIONS OF DISSEMBLING

Scale	N Condition		HIA Condition		D Condition	
	M	SD	M	SD	M	SD
<i>Tⁿ</i>	1.4	1.84	.3	.68	.2	.42
<i>L</i>	47.9	7.78	46.3	6.62	42.4	4.81
<i>Fⁿ</i>	3.1	2.96	6.0	3.83	25.0	11.87
<i>K</i>	58.1	8.49	51.3	8.79	42.9	8.99
<i>Hs</i>	52.4	7.28	50.3	5.16	70.8	15.61
<i>D</i>	47.8	5.41	56.1	12.50	79.3	17.31
<i>Hy</i>	56.2	7.73	54.9	9.42	71.0	7.09
<i>Pd</i>	53.8	10.90	59.0	13.26	80.5	14.49
<i>Mf</i>	42.2	8.16	45.5	9.16	49.7	10.24
<i>Pa</i>	49.4	6.13	60.4	10.85	87.3	21.49
<i>Pt</i>	54.1	7.95	61.4	12.77	85.4	14.91
<i>Sc</i>	55.2	6.48	65.3	10.87	98.8	19.25
<i>Ma</i>	60.6	10.30	63.7	11.85	71.5	11.44
<i>Si</i>	49.9	8.67	56.5	12.42	72.2	13.12
<i>A</i>	44.3	4.83	54.5	10.70	72.6	9.36
<i>R</i>	47.4	7.07	46.7	5.54	49.0	8.62

Note.—*N* = 10.

* Based on raw scores.

TABLE 2
ANALYSES OF SCORES UNDER NORMAL, HYPNOTICALLY INDUCED ANXIETY, AND
DISSEMBLING CONDITIONS

Scale	F^a	N and HIA		HIA and D		N and D	
		MD	t	MD	t	MD	t
P^b	3.31	-1.1	1.88	-.1	1.00	-1.2	2.09
L	1.88	-1.6	1.14	-3.9	2.03	-5.5	2.84
F^b	29.39**	2.9	3.65**	19.0	4.80**	21.9	5.50**
K	7.55**	-6.8	4.37**	-8.4	3.03*	-15.2	6.15**
Hs	11.80**	-2.1	1.60	20.5	4.22**	18.4	3.84**
D	16.48**	8.2	2.21	23.2	3.65**	31.5	5.89**
Hy	12.08**	-1.3	.79	16.1	4.23**	14.8	4.46**
Pd	11.91**	5.2	1.07	21.5	5.55**	26.7	4.20**
Mf	1.66	3.3	2.36	4.2	1.10	7.5	1.99
Pa	16.45**	11.0	2.78*	26.9	6.00**	37.9	6.04**
Pt	17.93**	7.3	2.17	24.0	5.94**	31.3	7.21**
Sc	29.44**	10.1	3.07*	33.5	6.13**	43.6	7.13**
Ma	2.51	3.1	1.45	7.8	1.82	10.9	2.47
Si	9.80**	8.6	4.00**	15.7	3.06*	22.3	4.42**
A	27.34**	10.2	4.31**	18.1	5.44**	28.3	10.90**
R	0.27	-.7	.56	2.3	.72	1.6	.45

Note.—Minus signs indicate that the scores for the second condition listed were lower than those of the first.

^a Values of F obtained by analysis of variance for each scale under the three conditions of the experiment.

^b Based on raw scores.

* Significant at the .05 level.

** Significant at the .01 level.

to create the test profile of a person suffering great anxiety.

In order to provide an additional subjective check of the subjects' emotional states during the experiment, an anxiety rating scale was constructed according to the Likert technique (Edwards, 1957). It consisted of 40 questions about the way the subject felt at the time of responding to the scale. It included items such as: "I am at ease," "I feel tense without any good reason," "My morale is low," "I am restless and irritable now." Each of its 40 items had been shown to discriminate between high anxious and low anxious groups. It has a split-half reliability coefficient of .97. This scale was administered with the MMPI as part of a classroom demonstration. It was also given under conditions of hypnotically induced anxiety.

RESULTS

The data of this experiment consisted of the MMPI profiles obtained under normal conditions (N), conditions of hypnotically induced anxiety (HIA), and conditions of dissembling (D), as well as scores on the anxiety rating questionnaire obtained under the first two conditions. The means and the standard deviations of the T scores for each scale are listed in Table 1. Both the F and F scales are given in raw scores. The scores on F were well enough below the necessary

30 that conversion to T scores would necessitate each raw score having the same T , i.e., a T of 50. The scores on F were so high under conditions of dissembling that conversion to T scores tended to hide differences between this condition and the other two. Because of the small N ($N = 10$), $N-1$ was used in computing the standard deviations (Edwards, 1950).

An analysis of variance was performed for each scale under the three conditions. A comparison was then made between the T scores on each scale obtained under N and HIA conditions, N and D conditions, and HIA and D conditions. A t test was employed for this purpose. Because the scores obtained under these three conditions were not random with respect to each other, a t comparing the differences between correlated means was computed using the differences between the scores (McNemar, 1949). The mean differences and the t 's for each scale for all combinations of the three conditions are listed in Table 2. Although all values of t were reported they were marked as significant, in the conventional manner, only for those scales where significantly large F s were obtained.

In comparing the scores between the N and HIA conditions, the *F*, *K*, *Pa*, *Sc*, *Si*, and *A* scales showed significant changes. In comparing the scores between the HIA and D conditions, only the *?*, *L*, *Mf*, *Ma*, and *R* scales showed insignificant changes. Likewise, the differences between the scores for N and D conditions were insignificant for the *L*, *Ma*, *?*, *Mf*, and *R* scales, being significant for all others.

A *t* was also computed for the anxiety rating questionnaire. The scores on this questionnaire, given twice (N and HIA conditions) showed a mean change which is significant at the .01 level.

DISCUSSION

Observation of Subjects

During the HIA session, there were indications of stress on the part of the subjects. When asked how they felt, they made comments such as: "I don't feel good," "I feel like I want to get out of here," "I feel unhappy," and "I feel as if something were wrong." In addition to these comments, the subjects showed signs that the experimenters interpreted as discomfort and distress. These signs were: furrowing of the brow, clenching of hands, frowning, tenseness, biting of lips, sighing deeply. One subject, a female, burst out crying while answering the MMPI items. The experimenter stopped the test and, seeing that she could not continue because of excessive crying, read the alleviating instructions. She stopped crying as the instructions were being read and agreed to complete the test the following day under the same experimental conditions.

In general, the subjects concentrated on the test and appeared to be making an effort to read and answer the items carefully. All subjects appeared relieved when the alleviating instructions were read. This relief was evidenced by smiling, relaxation of facial muscles, restrained laughter, and remarks such as: "I feel good now."

The A Scale

Hypothesis 1 stated that there would be a significant increase in the *A* scale between scores obtained under normal conditions and under conditions of hypnotically induced

anxiety. This hypothesis was supported by the data, the difference in the scores being significant at the .01 level.

The *A* scale is made up of items which occur on several of the other scales. Cluster and factor analyses indicated that the scale was relatively homogeneous and seemed to be related to anxiety. The scale contains very few "obvious" items that deal directly with the word anxiety and its synonyms. However, the experimenters found seven such items that they considered obvious with respect to the "anxiety" instructions the subjects received: "I feel anxiety about something or someone almost all of the time," "I must admit that I have at times been worried beyond reason over something that really did not matter," "I worry quite a bit over possible misfortunes," "I brood a great deal," "I wish I could be as happy as others seem to be," "Most of the time I feel blue," "I very seldom have spells of the blues." A count was made of the number of times these seven items were chosen under the three conditions, N, HIA, and D. They were chosen a total of 13 times under N condition, 34 times under HIA condition, and 63 times under D condition. These differences were significant at the .01 level. A *t* was then computed for the *A* scale with these seven items omitted to find if significant differences would still be obtained without the obvious items. The removal of these items did not alter the degrees of significance obtained previously with the full scale. This reduces the likelihood that the elevation of the *A* scale was the simple result of a heightened and conscious intent to comply with the suggestions of the experimenters.

The Validity Scales

Hypothesis 2 stated that the validity scales would differentiate between profiles obtained under conditions of dissembling and profiles obtained under conditions of both normal and hypnotically induced anxiety. This hypothesis was supported. Using the validity scales separately and the *F* minus *K* dissimulation index with a cutoff point of plus five (Gough, 1950), all profiles from both the N and HIA conditions were in the normal range indicating valid profiles. In addition, the *F* scale alone identified all but one of the

profiles obtained under conditions of dissembling. The *F* minus *K* index also did not identify this one profile, but identified all others.

The significant changes in the *F* and *K* scales obtained from Condition N to Condition HIA do indicate that a change in test-taking attitude occurred in the latter state. The *K* scores were significantly lower in Condition HIA as compared with Condition N, indicating that the subjects became more critical of themselves. In addition, the *F* scores were significantly higher in the HIA condition, indicating that while in this state, the subjects answered more of these items in the direction away from the direction the normal standardization groups answered them. The experimenters feel, however, that this might be expected in that the HIA state represents a condition removed from the condition under which standardization was obtained.

It is interesting to note that, although not significantly so, *L* scores were consistently lower in the HIA and D states as compared to N. Perhaps the criticalness of the subjects, as evidenced by their lower *K* scores in the HIA and D conditions, also made them more "honest."

An hypnotic scale has not been derived from the MMPI item pool. Admittedly such a scale would be of little clinical value, but it would be of considerable experimental interest. A scale capable of differentiating between the waking and the hypnotic state would serve as an objective device for indicating achievement of the trance state. It might also provide the basis for an objective method of appraising depth of trance.

The Diagnostic Scales

All diagnostic scales except *Mf*, *Ma*, and *R* showed significant differences from N to D conditions. All diagnostic scales except *Mf*, *Ma*, and *R* showed significant differences from HIA to D conditions. Since the validity scales identified dissembling profiles in Condition D, but not in Condition HIA, these differences further support these results in indicating a real difference between the two conditions. Profiles obtained under conditions of hypnotically induced anxiety do not resemble those obtained under conditions of

dissembling. The differences between profiles obtained in these two conditions suggest that hypnosis is not a state of mere heightened cooperation.

In going from Conditions N to HIA, it might be expected that only the *A* scale would be increased, since the instructions under hypnosis were directed to this effect. However, the *Pa*, *Sc*, and *Si* scales were also significantly heightened.

The rise in *Pa* may be considered as a direct result of the anxiety instructions. Looking post priori, it can be seen that suggestions such as "You are beginning to feel very uneasy and anxious," and "You don't know why, but this uneasy feeling is making you nervous, irritable, and frightened," might easily be reflected in the items composing the *Pa* scale as these items were derived from patient samples "symptomatically . . . to have ideas of reference, to feel that they were persecuted by individuals or groups. . . ." (Hathaway, 1956, pp. 109-110).

The heightened *Sc* and *Si* scales can be considered in the same way. Feelings of apprehension and anxiety, as well as feelings of being blue and melancholy, might be reflected in items dealing with social introversion, and schizophrenic symptoms have been described in much the same way. The *D* scale, however, was not significantly increased, indicating that the part of the instructions relating to feelings of depression were not alone responsible for these changes.

The change in scores in the Likert-type anxiety questionnaire from Conditions N to HIA were significant at the .01 level. The questionnaire was not given under Condition D because of its transparency. The change indicates that each subject's subjective evaluation of the way he felt during the HIA session corresponded to the overt behavioral differences observed and to the detection of these differences effected by the *A* scale. Most of the scores doubled and some increased by as much as 100 out of a possible 160 points. Only 1 subject out of 10 failed to report an increase in anxiety as reflected in the questionnaire.

It is impossible to estimate the effect of the order of the experimental conditions upon test performance. The order used in this design was chosen so as to minimize con-

tamination of experimental conditions by prior experience. Obviously the "normal" administration of the MMPI which served as a control and also as the basis for selecting subjects had to come first. It was felt that the faking conditions should be last in order to prevent the establishment of a faking set which might persist and intrude upon the hypnotically induced anxiety condition.

SUMMARY AND CONCLUSIONS

Ten college students took the MMPI under three conditions. In the first condition, the test was taken as part of a classroom demonstration. The second administration occurred under conditions of hypnotically induced anxiety. The third administration occurred in the waking state after instructions were given to "fake" anxiety.

Comparisons of the test profiles and the results of the specially constructed anxiety questionnaire permitted the following conclusions to be drawn:

1. Anxiety suggestions, when culled from definitions and descriptions of anxiety from independent sources, cause a significant increase in the Welsh A scale of the MMPI when they are given to hypnotized subjects.

2. Overt behavioral signs indicate that affective changes are experienced when anxiety is suggested to hypnotized subjects.

3. The anxiety questionnaire revealed that the subjects reported a marked increase of feelings of tension, discomfort, unpleasantness, and apprehension following the "anxiety" instructions in the hypnotized state over their reports in the normal waking state.

4. The validity scales of the MMPI successfully identify 9 out of 10 dissemblers and show that, in a state of hypnotically induced anxiety, valid profiles are obtained.

5. Significant differences in the diagnostic and validity scales between conditions of hypnotically induced anxiety and conditions of dissembling indicate that the former is different enough from the latter to strongly suggest that hypnosis is not a state of mere exaggerated cooperation.

REFERENCES

- COFER, C. N., CHANCE, J., & JUDSON, A. J. A study of malingering on the Minnesota Multiphasic Personality Inventory. *J. Psychol.*, 1949, 27, 491-499.
- CONKLIN, E. *Principles of abnormal psychology*. New York: Holt, 1936.
- DORCUS, R. M. (Ed.) *Hypnosis and its therapeutic applications*. New York: McGraw-Hill, 1956.
- EDWARDS, A. L. *Experimental design in psychological research*. New York: Rinehart, 1950.
- EDWARDS, A. L. *Techniques of attitude scale construction*. New York: Appleton-Century-Crofts, 1957.
- FISHER, V. E., & MARROW, A. J. Experimental study of moods. *Charact. Pers.*, 1934, 2, 201-208.
- GOUGH, H. G. Simulated patterns on the Minnesota Multiphasic Personality Inventory. *J. abnorm. soc. Psychol.*, 1947, 42, 215-225.
- GOUGH, H. G. The F minus K dissimulation index for the Minnesota Multiphasic Personality Inventory. *J. consult. Psychol.*, 1950, 14, 408-413.
- GROSZ, H. J., & LEVITT, E. E. The effects of hypnotically induced anxiety on the Manifest Anxiety Scale and the Barron ego-strength scale. *J. abnorm. soc. Psychol.*, 1959, 59, 281-283.
- HATHAWAY, S. R. Scales 5 (masculinity-femininity), 6 (paranoia) and 8 (schizophrenia). In A. S. Welsh & W. G. Dahlstrom (Eds.), *Basic readings on the MMPI in psychology and medicine*. Minneapolis: Univer. Minnesota Press, 1956.
- HEYNS, R. W. *The psychology of personal adjustment*. New York: Dryden, 1958.
- HUNT, H. F. The effect of deliberate deception on Minnesota Multiphasic Personality Inventory profiles. *J. consult. Psychol.*, 1948, 12, 396-402.
- HUSTON, P. E., SHAKOW, D., & ERICKSON, M. H. A study of hypnotically induced complexes by means of the Luria technique. *J. gen. Psychol.*, 1934, 11, 650-697.
- LEHNER, G. F., & KUBE, E. *The dynamics of personal adjustment*. New Jersey: Prentice-Hall, 1955.
- LURIA, A. R. *The nature of human conflict*. New York: Liveright, 1932.
- MCMENAR, Q. *Psychological statistics*. New York: Wiley, 1949.
- MAY, R. *The meaning of anxiety*. New York: Ronald, 1950.
- SHAFFER L., & SHOEN, E. *The psychology of adjustment*. (2nd ed.) Boston: Houghton Mifflin, 1956.
- SWEETLAND, A. Hypnotic neuroses: Hypochondriasis and depression. *J. gen. Psychol.*, 1948, 39, 91-105.
- WARREN, H. C. *Dictionary of psychology*. Boston: Houghton Mifflin, 1934.
- WELSH, G. S. An anxiety index and an internalization ratio for the MMPI. *J. consult. Psychol.*, 1952, 16, 65-72.

(Received March 28, 1960)

THE EFFECT OF BRAIN DAMAGE UPON SPEED, ACCURACY, AND IMPROVEMENT IN VISUAL MOTOR FUNCTIONING

KENNETH B. STEIN¹

Veterans Administration Regional Office, San Francisco

The purpose of this study was to investigate three aspects of visual motor functioning in relation to cortical deficit. These aspects were visual motor *speed*, *accuracy* in visual motor reproduction, and *improvement*. These features were observed simultaneously within a single unitary time span and a single unitary task. The relative importance of each of these modalities as well as their interrelationships were also studied.

Klebanoff (1945) and Klebanoff, Singer, and Wilensky (1954) have extensively reviewed studies of brain damage. The reviews reveal that some of these three visual motor features have been studied either implicitly or explicitly.

The speed variable is dealt with in timed tasks. Some of the performance subtests of the WAIS (Wechsler, 1955) lend themselves to measures or scores based on time or time bonus credits. Almost all investigators have found some of these visual motor subtests impaired by cortical damage such as the Digit Symbol (Klebanoff et al., 1954, p. 13) and the Block Design (Aita, Armitage, Reitan, & Rabinowitz, 1947; Allen, 1947, 1948; Goldman, Greenblatt, & Coon, 1946; Lidz, Greenblatt, Goldman, & Coon, 1946; Lidz, Gay, & Tietze, 1942). Another example of visual motor speed tasks was the cancellation test techniques used by Hunt (Freeman & Watts, 1939) and Rylander (1939). They found reduced speed following lobotomy. Each of them also employed an accuracy

score. Hunt found that subjects demonstrated greater accuracy after lobotomy. Rylander did not find any significant difference in accuracy scores.

As with tests for speed, there are specially devised techniques which focus upon the qualitative or accuracy aspect of functioning by the brain injured. The Bender-Gestalt (Bender, 1933) and the St. Louis Memory-for-Design (Graham & Kendall, 1946) Tests involve the reproduction of geometrical figures. In these tests it is often noted that subjects with cortical lesions tend to reorganize the figures qualitatively in the direction of simplification. Fragmentation, rotation, reversal, and greater closure and balance are some of the manifestations of simplification. These perceptual motor disturbances have been discussed in terms of Goldstein's theory of concrete and abstract attitudes (Goldstein & Scheerer, 1941) as well as gestalt theory (Bender, 1933).

One would expect that the concrete attitude of the organic would be revealed not only in its effect upon speed and accuracy but also upon improvement. Goldstein and Scheerer (1941) indicate that the organic does not profit as readily from experience as the normal. McFie and Piercy (1952) found impairment of retention and learning which was related to the size of the lesion.

The main hypothesis for investigation in the present experiment was that organics differ significantly from nonorganics in speed, accuracy, and improvement in visual motor functioning. As an auxiliary problem an exploration was made of the interrelationships of these three variables as well as two

¹ The author wishes to thank R. C. Tryon, University of California, Berkeley, and F. Henderson and staff members of the Veterans Administration Clinic, San Francisco, for their helpful suggestions and critical reading of the manuscript.

additional ones, IQ and age. Finally, each variable as well as the combination of these variables was assessed for power of discrimination between organics and nonorganics.

METHOD

Subjects

There were 60 subjects with cortical brain damage and 120 controls. All were white and American born. Of the 120 controls 15 were female and 105 male, whereas all the organics were male.² Forty-two of the experimental subjects were VA outpatients and the remaining 18 were hospitalized veterans. Diagnostically the organic subjects were distributed as follows: 27 posttraumatic encephalopathies, 14 cortical insults associated with circulatory disorders, and 8 preoperative and postoperative tumors. The remaining 11 experimental subjects had varying diagnoses such as paresis, encephalitis, tuberculous meningitis, and cortical atrophy. None of them was considered psychotic. Of the 120 controls, 51 were VA outpatients with nonpsychotic psychiatric diagnoses; 29, outpatients in treatment for tuberculosis; and 40, general medical outpatients. None of the experimental and control subjects had any noticeable visual defect or motor impairment of their writing hand.

The 180 subjects were drawn from a larger pool of 261 subjects in order that the organics and controls could be equated for age, education, and IQ. Other clinical groups such as various psychotics will be used in a subsequent study and should provide a comparison with the results on the present groups.

Procedures

Vocabulary Subtest. The Vocabulary subtest of the Wechsler-Bellevue scale, Form I (Wechsler, 1944) was used to obtain an estimate of verbal IQ. Morrow and Marks (1955) found no significant difference between brain injured and control subjects on vocabulary. This lack of difference led the authors to conclude that this measure was an adequate indicator of premorbid IQ. Both Jackson (1955) and Rapaport (1945) report that the Vocabulary subtest tends to be relatively refractory to impairment.

Symbol-Gestalt Test. It was expected that the organics with their concrete attitude would show significantly greater difficulty than nonorganics with both speed and accuracy as well as with improvement. As a means of tapping these three aspects of visual motor functioning, a symbol substitution

task was developed. The format is similar to the Digit Symbol subtest of the WAIS but new symbols have been devised³ (see Figure 1). These symbols were constructed to have poor gestalt form, i.e., they lack closure and balance. They have gaps and unequal as well as nonparallel lines. This visual motor task has a 3-minute time limit and the number of substitutions made for each minute was recorded by the experimenter. There are a total of 110 substitution items. The time limit prevents subjects from completing all items so that no one achieves a maximum score.

1	2	3	4	5	6	7	8	9
.	N	⊖	1	0	≡	7	⊗	=

FIG. 1. Symbol-Gestalt Test—symbols and numbers.

This task yielded one speed, one improvement, and three accuracy measures which could be statistically compared for the experimental and control groups. The speed score (3-minute complete) was the total number of substitutions made in the 3-minute time limit. The improvement measure was calculated by subtracting the total number of correct symbols in the first minute from the total number in the third minute. The three accuracy measures were: (a) the total number of correct substitutions in 3 minutes (3-minute correct); (b) the percent error (% error) based on the number of errors in the first 40 substitutions; and (c) the number of qualitative errors (Q error)—rotations, reversals, wrong substitutions, and distortions. The instruction for administration of the test to the subjects was similar to that of the Digit Symbol subtest (Wechsler, 1955).

RESULTS

Equality in Education, Age, and IQ

In the statistical treatment of the measures, *t* tests were calculated. Table 1 reveals first that the two groups were equated for education, age, and IQ. Since these variables may be related to speed, accuracy, and improvement independently of brain damage, it was necessary to hold them constant. Thus it was possible to assess the unique effects of organic injury upon the visual motor functions.

Differences between Organics and Controls

The controls demonstrated significantly greater speed than the organics as shown in the 3-minute complete score. The mean dif-

²As part of an earlier unpublished study the author found no significant differences for the sex variable on the Symbol-Gestalt procedure used in this experiment. The 22 females' mean score = .687, *SD* .536; 22 males' mean score = .606, *SD* .448. The *t* = .534 which is insignificant. Both groups were equated for age and IQ.

³Although this test was devised by the author, it was first used by Phelps (1952) in a modified form.

TABLE 1

MEANS, STANDARD DEVIATIONS, RANGES, AND *t* VALUES FOR 120 CONTROLS AND 60 ORGANICS

Variable	M_c	M_o	$M_c - M_o$	Range _c	Range _o	SD_c	SD_o	<i>t</i>
Education	11.17	11.23	.06	5-16	5-16	2.66	2.78	.003
Age	39.91	40.42	.51	20-67	21-65	14.08	13.68	.233
IQ	114.39	113.27	1.12	91-133	92-139	9.0	10.05	.732
3-minute complete	59.02	42.03	16.99	25-105	17-90	16.72	14.19	7.127**
3-minute correct	50.08	29.92	20.16	21-94	4-59	15.20	13.18	9.186**
% error	14.95	29.88	14.93	0-55	0-85	12.55	17.53	5.888**
Q error	.63	1.62	.99	0-15	0-13	1.59	2.56	2.729*
Improvement	2.16	.45	1.71	-5-+9	-9-+9	3.13	3.35	3.295**

* *p* at .01 level is 2.645.** *p* < .001.

ference of 16.99 symbols achieved a *p* at below the .001 level (see Table 1).

Of the three accuracy measures, the 3-minute correct score yielded the largest *t* value of 9.19, which has a *p* also below the .001 level. The controls completed 20.16 more correct symbols than the brain damaged subjects.

The % error, the second accuracy measure, revealed that the controls committed half as many errors as the experimental group, resulting in a *t* of 5.89 which is significant at less than the .001 level. The organics had a wider range and greater variability of % error scores.

In the Q error, the third accuracy measure, the organics produced almost three times as many such errors as the nonimpaired group. The *t* of 2.73 is significant at less than the .01 level. Although the controls showed a wider range of these errors (0-15) than the organics (0-13), this was due to one control subject who was the only one with

more than 5 such errors. Over 10% of the organics made 5-13 qualitative errors.

Improvement, the last measure, showed the organics had improved very little whereas the controls had a mean improvement of more than two symbols. The *t* value was 3.29 which is significant at the .001 level.

Relationship between Variables

To ascertain a measure of amount of impairment of the three aspects of visual motor functioning, point biserial correlations were calculated between each score and the dichotomous criterion variable, i.e., organic-nonorganic. In order of size the correlations for the five scores were as follows: 3-minute correct .548, 3-minute complete .451, % error -.440, improvement .245, and Q error -.230. In addition age and IQ had near zero correlations, .018 and .056, respectively.

Table 2 shows the intercorrelations for the seven variables on the entire sample. Although there were a number of significant

TABLE 2
INTERCORRELATIONS FOR TOTAL SAMPLE ON SEVEN VARIABLES

Variable	IQ	3-Minute Complete	3-Minute Correct	% Error	Q Error	Improvement
Age						
IQ	—					
3-minute complete	-.093	—				
3-minute correct		-.549	—			
% error		.248	-.585	—		
Q error			.173	.267	—	
Improvement			.895	.077	.182	—
				-.196	.068	-.147
				-.581	-.117	.085
					.555	.203
						.233
						-.085
						-.066

Note.—Correlations of .193 and .146 are significant at the .01 and .05 levels, respectively.

TABLE 3
INTERCORRELATIONS FOR CONTROL AND ORGANIC GROUPS ON SEVEN VARIABLES

Variable	IQ		3-Minute Complete		3-Minute Correct		% Error		Q Error		Improvement	
	C	O	C	O	C	O	C	O	C	O	C	O
Age	-.244	.187	-.602	-.620	-.680	-.708	.235	.379	.140	.250	-.193	-.059
IQ			.323	.101	.287	-.066	.008	.251	.025	.152	.048	.116
3-minute complete					.874	.819	.093	-.151	.072	-.149	.069	.195
3-minute correct							-.374	-.616	-.192	-.380	.109	.155
% error									.458	.576	.020	.035
Q error											-.017	.005
Improvement												

Note.—For controls .234 and .179 are significant at .01 and .05 levels, respectively; for organics .328 and .252 are significant at .01 and .05 levels, respectively.

correlations, only a few were large enough to account for a sizeable amount of the variance. The speed score (3-minute complete) correlated .895 with one of the accuracy scores, 3-minute correct, but showed low correlations with the remaining two accuracy scores (with % error $-.196$, with Q error $-.117$). The speed score also had a high relationship of $-.549$ with age. Likewise the 3-minute correct accuracy score was highly related to age ($-.585$).

As might be expected the three accuracy measures were significantly correlated with each other. The 3-minute correct with % error yielded an r of $-.581$, 3-minute correct with Q error was $-.335$, and Q error with % error was $.555$. Thus 3-minute correct correlated negatively with the other two accuracy scores, whereas Q error and % error were positively related to each other.

In order to compare the organic and control groups further, a correlation matrix

was obtained for each group on the seven variables. These correlations are found in Table 3. There were a few correlations which seem to be divergent for the two groups. The correlations between age and IQ went in opposite directions for the two groups. The controls showed an r of $-.244$ which is significant at the .01 level. The organics showed an r of $.187$ which falls short of significance. Yet the difference between these two correlations transformed to Fisher's z' values yielded a z equal to 2.284 which is significant at less than the .05 level. The nonorganics showed a significant relationship between IQ and 3-minute correct (.295) but the organics did not ($-.066$). The difference is significant with a z of 2.228 .

The organics had a higher correlation between 3-minute correct and % error ($-.616$) than the controls ($-.374$). The z of 2.006 is again significant for the difference be-

TABLE 4
THE CLASSIFICATION ACCURACY OF EACH VARIABLE BASED ON 120
CONTROLS AND 60 ORGANICS

Variable	Cutoff Score	# Overlap			% Overlap			% Correct Classification
		C	O	T	C	O	T	
3-minute complete	42	28	18	46	23	30	26	74
3-minute correct	37	17	20	37	14	33	21	79
% error	31	11	33	44	9	55	24	76
Q error	3	5	48	53	4	80	29	71
Improvement	0	32	34	37	27	57	37	63

TABLE 5

THE CLASSIFICATION ACCURACY FOR SIX VARIABLES
COMBINED USING *B* COEFFICIENTS

Group	N	Over- lap	% Over- lap	% Cor- rect Classi- fication
Controls	120	13	10.83	89.17
Organics	60	8	13.3	86.7
Total	180	21	11.67	88.33

tween the two correlations. The remaining correlational differences are not significant.

Diagnostic Classificatory Power of the Measures

Table 4 indicates the empirical cutoff point which maximizes differentiation between organic and nonorganic groups for each variable. These cutoff points were determined on the present sample. The 3-minute correct score had the highest discrimination with 79% correct classification. Next was % error with 76% followed by 3-minute complete, *Q* error, and improvement. In each case the controls displayed a smaller percentage of misclassification than did the organics.

The classificatory power of these variables taken singly and in combination is shown in the comparison of Tables 4 and 5. In a previous study involving 261 subjects from which the present 180 subjects were drawn, the *B* coefficients were determined for each of these five variables as well as for the sixth, i.e., age.⁴ These *B* coefficients were applied to the subjects in the current study. Therefore the results are not strictly cross-validative in nature. The results reflected a much smaller misclassification for both groups. When the six variables were combined, the controls showed a misclassification of 10.83% and the organics 13.3%. For the total sample the incorrect classification was 11.67%.

DISCUSSION

The results indicate that the organics are significantly impaired in speed, accuracy, and

improvement in visual motor functioning. Since the r_{pb} of the 3-minute correct is of greater magnitude than that of the other scores, it suggests that the organics suffer most in the accuracy function. Further investigation raises some question about such a conclusion. Table 2 discloses an extremely high correlation of the 3-minute correct accuracy measure with the speed score indicating that the 3-minute correct is not strictly an independent accuracy measure. Since this score is also highly correlated with the other two accuracy measures, it suggests that the 3-minute correct taps a combination of the speed and accuracy functions. Both the *Q* error and % error with their low correlations with speed appear to be fairly independent of speed. In view of these findings, a reconsideration of the hierarchy of r_{pb} 's suggests that the speed function is most impaired, closely followed by one of the accuracy measures, i.e., % error, and then by improvement.

Since the 3-minute correct score seems to be a combination of both speed and accuracy as well as having the largest r_{pb} , it suggests that a combination of variables or functions is more likely to reveal a greater difference between groups than any single variable. Evidence for this inference is pointed up in a comparison of Tables 4 and 5. Not only is the percent correct classification greater for the combined variables, but also the organic and control groups show less divergence since they are within 2.5% of each other.

An additional reason for the higher discriminatory power of the combined scores is the inclusion of the age variable. Since the two groups were equated for age, this variable had no effect by itself in separating organics from nonorganics. Yet age did have a definite influence within groups upon visual motor performance since it was significantly correlated with the three main variables. These correlations reveal that as age increases, speed, improvement, and accuracy decrease. These findings are similar to Wechsler's results (1944, 1955). The function that age serves in combination with the other scores is that of a suppressor variable.

The IQ variable shows a low but significant correlation with speed but no relationship to accuracy and improvement. These findings

⁴ An unpublished preliminary study by the author. *B* coefficients were as follows: 3-minute correct .00779, 3-minute complete .02194, age .02087, % error -.02068, *Q* error -.00624, and improvement .04118.

are at variance with Wechsler (1955) who revealed a correlation of .64 between Vocabulary IQ and the Digit Symbol subtest. This difference can be explained partly by the absence of the lower quartile from the IQ range in the present study. Another factor which may contribute to the lower correlation is that organics with negligible correlations are more heavily weighted in the present sample than would be expected in the general population with which Wechsler dealt.

The results in Table 3 show three sets of correlations to be significantly different for the two groups. On age with IQ the two groups went in opposite directions. The explanation may be that merely by chance this organic group was more heavily weighted for the higher IQs in the older age range than the controls. Likewise by chance the controls were more heavily weighted for lower IQs in the younger age range. The correlation of IQ and 3-minute correct shows the controls tending in the direction of Wechsler's result (1955) of a definite influence of IQ upon the Digit Symbol subtest score. With the impairment in visual motor functioning in the organic, IQ seemingly plays very little part in affecting the magnitude of his score. The third set of correlations that yielded a significant difference between the two groups was 3-minute correct and % error. The factor which may account for the difference is that the organic group made a significantly greater number of errors thus lowering their 3-minute correct scores.

If we exclude the 3-minute correct score, which was found to be a combination of speed and accuracy, Table 2 presents the striking finding that the intercorrelations are low for the measures tapping the three functions of speed, accuracy, and improvement. This suggests that the three functions are relatively independent. The conclusion of independence, however, has to be made with certain reservations. Since all the scores are derived from a single measuring instrument administered on one occasion, they are based to some extent upon responding to the same items. The speed score is based on all the correct and incorrect symbol substitutions. Therefore, error scores involve an overlap of certain items contained in the speed score. Similarly, the improve-

ment score has components of both the speed and accuracy measures. In order to test further for the independence of these three functions, a design in which measures are derived from experimentally different items will be required.

It remains for future studies to determine how the current results compare with samples involving other clinical groups such as schizophrenics and psychotic depressions. Confusion, motor retardation, and agitation in these groups may have disturbing effects upon the speed, accuracy, and improvement functions.

SUMMARY

Speed, accuracy, and improvement in visual motor functioning were investigated in 60 organic and 120 nonorganic subjects. A 3-minute substitution task was employed to obtain the data. This task yielded one speed, one improvement, and two strictly accuracy measures.

The results indicated that speed, accuracy, and improvement in visual motor performance were all significantly impaired in the brain injured group. The relatively low intercorrelations found suggested that the three variables may be fairly independent and specific factors contributing to the more general visual motor function or process. A fourth variable, age, appeared to be less independent since it correlated significantly with the first three variables. The fifth variable, IQ, unlike the first four, showed no noticeable influence upon visual motor performance. The implications of these findings were discussed.

The discriminant power of each of the scores was studied in relation to the number of correct classifications of organic and nonorganic subjects. These findings were then compared with the discriminant power of the combination of all of the scores. The result was that the combined scores yielded a higher percentage of correct classifications than the individual measures.

REFERENCES

- AITA, J. A., ARMITAGE, S. G., REITAN, R. M., & RABINOWITZ, A. The use of psychological tests in the evaluation of brain injury. *J. gen. Psychol.*, 1947, 37, 25-44.
- ALLEN, R. M. The test performance of the brain injured. *J. clin. Psychol.*, 1947, 3, 225-230.

- ALLEN, R. M. The test performance of the brain diseased. *J. clin. Psychol.*, 1948, 4, 281-284.
- BENDER, LAURETTA. Disturbances in visuomotor gestalt function in organic brain disease associated with sensory aphasia. *Arch. Neurol. Psychiat.*, 1933, 30, 514-537.
- FREEMAN, W., & WATTS, J. W. An interpretation of functions of the frontal lobe based upon observations in forty-eight cases of prefrontal lobotomy. *Yale J. Biol. Med.*, 1939, 11, 527-539.
- GOLDMAN, R., GREENBLATT, M., & COON, G. P. Use of the Bellevue-Wechsler scale in clinical psychiatry with particular reference to cases with brain damage. *J. nerv. ment. Dis.*, 1946, 104, 144-179.
- GOLDSTEIN, K., & SCHEERER, M. Abstract and concrete behavior: An experimental study with special tests. *Psychol. Monogr.*, 1941, 53(2, Whole No. 239).
- GRAHAM, FRANCES K., & KENDALL, BARBARA S. Performance of brain damaged cases on a Memory-For-Designs Test. *J. abnorm. soc. Psychol.*, 1946, 41, 303-314.
- GREENBLATT, M., GOLDMAN, R., & COON, G. P. Clinical implications of the Bellevue-Wechsler test (with particular reference to brain damage cases). *J. nerv. ment. Dis.*, 1946, 104, 438-442.
- JACKSON, C. V. Estimating impairment on Wechsler Bellevue subtests. *J. clin. Psychol.*, 1955, 11, 137-143.
- KLEBANOFF, S. G. Psychological changes in organic brain lesions and ablations. *Psychol. Bull.*, 1945, 42, 585-623.
- KLEBANOFF, S. G., SINGER, J. L., & WILENSKY, H. Psychological consequences of brain lesions and ablations. *Psychol. Bull.*, 1954, 51, 1-41.
- LIDZ, T., GAY, J. R., & TIETZE, C. Intelligence in cerebral deficit states and schizophrenia measured by Kohs Block Test. *Arch. Neurol. Psychiat.*, 1942, 48, 568-582.
- McFIE, J., & PIERCY, M. F. Intellectual impairment with localized cerebral lesions. *Brain*, 1952, 75, 292-311.
- MORROW, R. S., & MARKS, J. C. The correlation of intelligence and neurologic findings of twenty-two patients autopsied for brain damage. *J. consult. Psychol.*, 1955, 19, 283-289.
- PHELPS, C. K. An analysis of the integrative aspects of the performance of normal and brain damaged subjects on specially devised motor tests. Unpublished doctoral dissertation, Kansas University, 1952.
- RAPAPORT, D. *Diagnostic psychological testing*. Vol. I. Chicago: Year Book, 1945.
- RYLANDER, G. *Personality changes after operations on the frontal lobes: A clinical study of thirty-two cases*. Copenhagen: Ejnar Munksgard, 1939.
- WECHSLER, D. *Measurement of adult intelligence*. Baltimore: Williams & Wilkins, 1944.
- WECHSLER, D. *Wechsler Adult Intelligence Scale*. New York: Psychological Corp., 1955.

(Received March 28, 1960)

THE INFLUENCE OF CONTEXT ON THE DEPRESSION SCALE OF THE MMPI IN A PSYCHOTIC POPULATION

GORDON W. OLSON

Anoka State Hospital, Minnesota

This report is felt to be of current interest and value in view of the recent marketing of many antidepressant drugs, an influx which has created the need to objectively assess their effectiveness. Depression is often an especially difficult symptom to assess clinically because of the variety of manifestations and because it is frequently masked by or mixed with more dramatic symptoms such as schizophrenic apathy and hypochondriacal complaints. The present study was the result of a search for a short, reasonably objective, and easily procured measure of depression. A striking possibility appeared to be the depression scale (*D*) of the MMPI, the purpose of which was to identify the "state of mind characterized by poor morale, lack of hope in the future, and dissatisfaction with one's own status generally" (Hathaway & McKinley, 1942). The question posed was whether the *D* scale by itself would measure the same thing as when the entire inventory is administered, a question that is not answered by previous studies of reliability of the *D* scale.

Canter (1960) mentions the often-heard criticism that different response sets may be elicited if items or single scales are isolated from a main body of items, but found in his study that the *D* scale appearing out of context (but in combination with *Pt* and *K*) did differentiate among suicidal, nonsuicidal psychiatric, and nonhospitalized groups. This attests to the validity of the *D* scale and suggests that context played a not-too-important role in that instance.

METHOD

The 60 items comprising the *D* scale of the MMPI were mimeographed as a separate and columns provided to check the items as true or false. Fifty psy-

chiatric inpatients at Anoka State Hospital were administered consecutively the entire MMPI and the 60-item *D* scale. The sample contained 30 females and 20 males. Subjects (*Ss*) did not know beforehand that a second procedure would follow the first. One-half of the group took the entire MMPI first and the other half took the *D* scale only first. *Ss* were not selected with the exception that many had been referred for psychological examination and others were seen for routine testing on admission. Statistical analysis involved a *t* test of the difference for related means of the raw scores under the two conditions and the Pearsonian correlation coefficient for the 50 pairs of scores.

RESULTS AND DISCUSSION

The mean raw score when the *D* scale was administered alone was 22.1 and it was 22.1 when a part of the entire test. The standard deviations were 6.1 and 6.7, respectively. The *t* for this difference was .02, nonsignificant. The *r* for all pairs of scores was $.99 \pm .14$. The largest difference found in this sample was six points and the median difference was zero.

While the results of this study do not bear on the validity of the *D* scale as a measure of depression, they do clearly indicate that whatever is measured by the *D* scale of the MMPI can be measured in a psychiatric population without administration of the entire inventory.

SUMMARY

The recent influx of antidepressant drugs stimulated the search for a short, objective, and easily procured measure of depression. The *D* scale of the MMPI would seem to fulfill these criteria if it could be administered separately, but the question presented was the familiar one of effect of context on response set. Fifty psychiatric inpatients were individually administered the entire MMPI

and the *D* scale only; one-half took the MMPI first and one-half the *D* scale first. The correlation of *D* scores in the two situations was .99 and the *t* nonsignificant. It was concluded that context did not influence the response set and that whatever is measured by the *D* scale can be measured without administration of the entire inventory.

REFERENCES

- CANTER, A. The efficacy of a short form of the MMPI to evaluate depression and morale loss. *J. consult. Psychol.*, 1960, 24, 14-17.
- HATHAWAY, S. R., & MCKINLEY, J. C. A multiphasic personality schedule (Minnesota): III. The measurement of symptomatic depression. *J. Psychol.*, 1942, 10, 249-254.

(Received February 26, 1960)

SUPPRESSING DISTORTION IN TEMPERAMENT INVENTORIES

JACK HAND AND HERBERT H. REYNOLDS
Baylor University

This study was designed to determine the effects of "appraisal" and "research" testing conditions upon six scales taken from the Guilford-Zimmerman Temperament Survey (GZTS).

Subjects consisted of 373 USAF basic trainees divided into two groups. In the first week of duty both groups completed a self-report temperament inventory composed of 240 items from the GZTS. Instructions for the Research Group were:

I am studying personality tests and need your help. This work has no connection with the Air Force. Please do not put any identifying marks on your answer sheets.

This was in addition to the regular instructions. These instructions were administered by the senior author in civilian clothes. The Appraisal Group was told:

As you know, personality characteristics are related to success in the Air Force. The tests you are about to take will be a matter of record.

These instructions were administered by the co-author in full uniform (Air Force Captain).

Included in the inventory are the following scales:

DG—8 socially desirable items from the GZTS G scale¹

UG—8 socially undesirable items from the GZTS G scale

DR—10 socially desirable items from the GZTS R scale

UR—10 socially undesirable items from the GZTS R scale

SD—a scale designed to measure SD²

¹ SD values for all items in the GZTS (except MF) were established by another investigator (Kelley, 1959).

² A description of this scale is being prepared for publication. The correlation between it and Edwards SDS (Edwards, 1957) is .54.

TABLE 1
COMPARISON OF APPRAISAL AND RESEARCH GROUPS
ON SIX CLUSTERS OF ITEMS FROM THE GZTS

Variable	Appraisal (N = 190)		Research (N = 183)		<i>t</i> ratio (<i>M</i> ₁ - <i>M</i> ₂)
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	
DG	5.25	1.79	4.84	1.80	2.20*
DR	6.28	2.16	5.67	2.26	2.66**
UG	3.52	1.78	3.95	1.69	2.39*
UR	3.56	1.85	4.43	1.98	4.39**
DG + UG	8.77	2.89	8.79	2.76	.07
DR + UR	9.84	3.09	10.10	3.48	.76

* $p < .05$.

** $p < .01$.

Each subject received seven scores: DG, UG, DG + UG, DR, UR, DR + UR, and SD. The DG + UG and DR + UR are scores from scales composed of an equal number of desirable and undesirable items.

RESULTS AND DISCUSSION

Table 1 gives comparisons of group means on the temperament variables. These results indicate rather clearly the effects of different instructions and conditions upon the temperament scores. When defensiveness is stimulated the scores are increased. The insignificant group differences on variable DG + UG and DR + UR indicate, however, that a balanced design eliminates the effects of defensiveness (provided appraisal conditions stimulate defensiveness).

The product-moment correlations (for appraisal group) of SD with DG, DR, UG, UR, DG + UG, and DR + UR are .32, .49, -.34, -.46, .00, and .03 (first four significant at .01 level) further suggesting that the balanced design eliminates the influence of

SD upon temperament scores obtained under appraisal conditions.

Probably the most important interpretation of this data is that with a balanced design one set of norms may be appropriate for the two most widely applicable testing conditions.

REFERENCES

- EDWARDS, A. L. *The social desirability variable in personality assessment and research*. New York: Dryden, 1957.
- KELLEY, F. G. *Construction of a scale for the measurement of social desirability*. Unpublished master's thesis, Baylor University, 1959.

(Received March 18, 1960)

BRIEF REPORTS

SEDATIVES AND SUGGESTIBILITY IN NEUROTIC PATIENTS¹

J. G. INGHAM²

Llandough Hospital, Glamorgan, Wales

AND J. M. WHITE

Stanley Royd Hospital, Wakefield, Yorkshire, England

Ingham (1955) found that neurotic patients taking sedatives were significantly more suggestible than unsedated neurotics. Suggestibility might have been increased by sedation in these patients, or alternatively the more suggestible patients might, though not necessarily intentionally, have been selected for sedative treatment. Two investigations were done to find out which of these interpretations was more likely.

In the first investigation, 10 male neurotic patients were given the same medication (6 grains of sodium amytal per day, in divided doses) for 3 days following admission to hospital. They were then divided at random into two groups and took part in a 2-day experiment. One group was tested (a) after a further day on the same regime and again (b) after one day without drugs. For the other group, b was followed by a. The arm-movement test of suggestibility was used, as well as a test of static ataxia and arm-movement without suggestion.

This preliminary experiment offered no support for the idea that moderate sedation increases suggestibility but its value was limited by the small number of subjects and by the fact that the period without drugs was so short.

¹ An extended report of this study may be obtained without charge from J. G. Ingham (Medical Research Council Social Psychiatry Research Unit, Llandough Hospital; Penarth, Glamorgan, Wales; United Kingdom) or for a fee from the American Documentation Institute. Order Document No. 6543 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance \$1.25 for microfilm or \$1.25 for photocopies. Make checks payable to: Chief, Photoduplication Service, Library of Congress.

² Previously in the Medical Research Council Neuropsychiatric Research Unit, Whitchurch Hospital, Cardiff, Wales.

In a second investigation, an attempt was made to eliminate the difference in suggestibility between sedated and unsedated patients by retesting them after giving the same amount of sedatives to both groups for a few days. If the difference is a result of sedation, then it should be possible to eliminate it by such a procedure. If, on the other hand, the difference results from selection, then it should remain, even when the groups are equally sedated.

Forty-two male neurotic patients (15 of them sedated before admission) were tested on admission to hospital and again 3 days later. During the intervening period all patients received 6 grains of sodium amytal per day, in divided doses. It was again found that, tested on admission, previously sedated patients showed significantly greater arm-movement suggestibility than unsedated patients, though both were significantly more suggestible than a group of 27 normal males. On retest, after 3 days of medication, the significant difference in suggestibility between previously sedated and unsedated patients remained. There was no indication of an increase in suggestibility, following medication, in either group. The findings support the hypothesis that there was a selection factor operating, whereby the more suggestible patients were more likely to be receiving sedatives.

Additional results from both investigations suggest that static ataxia increases following moderate sedation.

REFERENCE

- INGHAM, J. G. Psychoneurosis and suggestibility. *J. abnorm. soc. Psychol.*, 1955, 51, 600-603.

(Received May 27, 1960)

SEMANTIC DIFFERENTIAL RATING OF SELF AND OF SELF-REPORTED PERSONAL CHARACTERISTICS¹

JAMES E. MADDEN

Veterans Administration Hospital, Chillicothe, Ohio

The semantic differential has been used as a measure of similarity of affective reactions to various concepts. When two or more concepts are rated similarly (have close profile agreement) the fact can be interpreted in clinical psychological terms. For example, equating the "self" concept and "descriptive" concepts suggests that the descriptive concepts are important aspects of the individual's sense of self. The present study reveals the extent to which this exemplified interpretation may be justifiable. The semantic differential profiles of "I, myself" and independently indicated aspects of self (self-reported personal characteristics) are compared, to see if in fact self and aspects are rated similarly. In addition to "close" profile agreement the study deals with other degrees of agreement, to show the relation between profile agreement and the probability that a descriptive concept is an aspect of self.

Fifty-nine concepts (personal characteristics) were derived from *Mf* scale items of the MMPI. For the rating task, the items were reworded and presented in the third person singular (i.e., "A person who . . ."). All items were rated by each *S*. However, only the concepts derived from items marked True by *S* during an independent administration of the items in their usual self-report form, are considered to represent aspects of his

sense of self. When *S* marks an item True, he virtually implies "I am a person who . . ." Also rated was the concept "I, myself."

A set of 15 bipolar seven-step scales (five evaluative, five potency, and five activity) was used to rate each concept. The square root of the sum of squared distances was employed as the measure of profile agreement or Distance between "I, myself" and each of the other concepts.

Thirty college students were *Ss*. The *Mf* items, in their usual True-False form, were administered to half the *Ss* approximately a week before they rated the concepts on the semantic differential. The remaining half had the reverse sequence of tasks.

The range of Distances for each *S* was divided into tenths. In each *S*'s data the percentage of True items (those which he had independently marked True) was computed in each of the 10 regions. Data were adjusted so that a 50% value for a region meant that True items had no greater or less tendency than False items did to occur in that region. The percentages of True items in the regions were then averaged across *Ss*.

For all 30 *Ss* the resulting percentage quantities, from Region 1 (closest to "I, myself") through Region 10 (most distant from "I, myself"), were as follows: 66.8, 65.1, 56.8, 51.6, 45.7, 40.4, 33.5, 23.9, 21.2, 20.6. The mean percentage of True items in the first five regions is significantly larger than the mean percentage in the second five regions, beyond the .01 level. Results were similar for sex and sequence groups of *Ss*. The probability that a descriptive concept is an aspect of self, is seen to be directly related to the amount of agreement between the ratings of self and the concept. From the data's negative facets (e.g., some False items in close regions and some True items in distant regions), some challenging theoretical possibilities can be gleaned.

(Received June 13, 1960)

¹ An extended report of this study may be obtained without charge from James E. Madden (Clinical Psychology Service, Veterans Administration Hospital; Chillicothe, Ohio) or for a fee from the American Documentation Institute. Order Document No. 6542 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance \$1.25 for microfilm or \$1.25 for photocopies. Make checks payable to: Chief, Photoduplication Service, Library of Congress.

The data reported here are from the writer's master's thesis completed at the University of Kentucky in 1956.

FIELD DEPENDENCE, MANIFEST ANXIETY, AND SOCIOMETRIC STATUS IN CHILDREN¹

IRA ISCOE AND JOYCE ANN CARDEN

University of Texas

The Children's Manifest Anxiety Scale (CMAS) as a measure of anxiety or drive level, a rating method of sociometric choice (rating own sex) employing three choice criteria as an index of social status, and Witkin's Embedded Figures Test as a measure of field dependence-independence were each administered to an entire sixth grade class composed of 16 boys and 15 girls. Mean age of the girls was 11 years 7 months, for the boys 11 years 11 months. Mean IQ was 118 for girls, 116 for boys. They were homogeneous with respect to religion and socioeconomic class. They had been acquainted for at least 7 months prior to the study.

A significant rank-order correlation of .57 ($p < .05$) was obtained between sociometric status and field dependence in girls. For boys the results were all in the opposite direction, with the choice of "class officer" correlating $-.51$ ($p < .05$) with field dependence. The data suggest that popular boys are more likely to exhibit an active field analytic orientation and girls a passive field dependent one. CMAS scores were not related to sociometric status for boys while significant negative correlations ($p < .05$) on all criteria questions were obtained for girls. In addition, the number of rejections received was significantly related to anxiety level, the r being .65 ($p < .01$). It would appear that more frequently chosen girls tend to have a lower drive level (anxiety) while under chosen girls have a higher level. The more rejected a girl is by her peers, the higher

her drive level. For boys, the correlation between drive level and field dependency was positive but not significant. For girls an r of $-.60$ was obtained ($p < .01$). This would indicate that field independent girls tend to be more anxious than field dependent ones. Level of intelligence was not significantly related to any of the sociometric choices, nor to scores on the CMAS. Some support for a significant negative relationship between field dependence and intelligence was obtained for girls but not for boys.

The results offer some support to the findings of Witkin and his associates, in regard to the field dependence-independence dimension and personality characteristics. In the present study, at the age and cultural level employed, the girl who is an active initiator and organizer is not likely to enjoy high social status with her peers. In contrast, the relatively field independent boy is most likely to gain wider acceptance by his classmates.

The descriptions Witkin uses to characterize field independent vs. dependent persons might well represent the kinds of behavior our middle class culture fosters and rewards at these ages. Boys are expected to be somewhat aggressive, direct, and analytic, while girls are taught a more submissive, conforming, "ladylike" type of behavior. The girl who identifies with this role gains acceptance and is subjectively aware of fewer discomforts as picked up by the CMAS. For girls, an analytic (field independent) mode of perceiving results in less popularity and more anxiety. The reversal of the relationship between field dependence and social status in boys perhaps emphasizes the cultural rewards for their exhibiting initiative. The fact that the types of behavior examined in the current study have some correlates to modes of perceiving points again to the intricate relationships that exist between perception and personality.

¹ An extended report of this study may be obtained without charge from Ira Iscoe (Department of Psychology, University of Texas; Austin 12, Texas) or for a fee from the American Documentation Institute. Order Document No. 6541 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance \$1.75 for microfilm or \$2.50 for photocopies. Make checks payable to: Chief, Photoduplication Service, Library of Congress.

(Received August 1, 1960)

THE ROLE OF THE INTERNSHIP IN THE RESEARCH TRAINING OF THE CLINICAL PSYCHOLOGIST

JULES D. HOLZBERG¹

Connecticut State Hospital

Training in clinical psychology has consisted of two aspects: (a) a minimum of 3 years at a university where the student receives a generic education in basic psychology, training in theory and techniques relevant to clinical psychology, and practicum experience in research; (b) at least 1 year of practicum experience (internship) in a clinical center to provide substance for the more academic aspects of the student's clinical training. Within this organization of clinical training, the university has been viewed as basically responsible for the training of the clinician as a psychological researcher, while the internship is generally conceived of providing only minimal support for the research aspects of a clinical student's training.

However, one may accept this basic responsibility of the university while at the same time questioning the adequacy of a conception that denies to the internship a significant responsibility in the training of the clinician as a researcher. The university is unquestionably better equipped from the point of view of facilities and personnel to provide training in the definition and application of psychological concepts, in the use of the library as a research tool, in experimental and research design, in the utilization of statistical analyses, in methods of sampling populations, and the numerous other requisite skills of the competent researcher.

However, the internship center has a significant role to play in the research training of the clinical student in at least two significant respects: (a) to sensitize the student to

the meaningful research problems in the clinical field and to demonstrate how clinical skills and techniques may be adapted to research goals, while providing a framework that fosters research on clinical problems without destroying their clinical significance; and (b) to contribute to the building of a self-concept that effectively integrates the role of "researcher" and "clinician." The first of these aspects is by no means insignificant, for the university may, because of its frequent insistence on the absolute "researchability" of a problem, force the clinical student into a research area where the problem is "neat," but the clinical significance quite meager. This is not meant as a criticism of university training, since the author recognizes that the function of the university is to provide generic training in the development of research skills, and there may be real merit in focusing this training on "researchable" problems. We must, however, be prepared for one possible consequence of this training, i.e., this criterion of researchability may become so rigidly fixed that the clinician may never undertake research on a clinical problem, because it is rare indeed that a clinical problem can hope to attain the level of researchability that meets the standard set for students in the typical university.

The second research training function of the internship is related to the author's observations that clinical interns frequently come to the internship with a strong sense of inferiority as a researcher. Whether this is a function of the preselection of clinical students (in that students entering into the clinical area may do so because of intellectual and personality reasons that are incompatible with a research role) or whether this is a

¹ The author is indebted to Solomon E. Feldman, who read the manuscript and helped make it a more readable document. However, the author assumes full responsibility for the ideas expressed.

function of the absorption of attitudes toward the clinician as a research worker that unfortunately permeate many of our universities, the result often seems to be that when the clinical student enters into the internship, he does so with a self-concept that is relatively weak with regard to the integration of research as a practicing role.

The internship frequently does not help to counter such feelings of inferiority. It may fail to provide a "culture" which rewards the development of a research role for the clinician by encouraging other role behaviors, such as that of diagnostician and therapist. The practicing clinician in the internship center may further contribute to this self-concept by failing to provide an adequate model of the clinician as a research worker. He is usually not involved in research and may actually give voice to antiresearch attitudes, so that we have the vicious "cultural" cycle of the clinician, having learned to reject the role of research for himself, now fostering the same attitudes in the students sent to him for training.

THE ACQUISITION OF A RESEARCH ROLE²

I have been using two concepts which should be more precisely defined. One is that of "role" of a researcher, i.e., those actions which are performed or not performed to confirm the occupancy of the research role. The second, "self-concept," may be defined as the values used by a person to describe and evaluate himself, and here I am specifically referring to self-descriptions of the clinician with regard to values pertinent to being a research worker. These two concepts are functionally quite interrelated since attitudes toward the self affect the kind of role that will and can be played, and conversely, the success with which one plays or does not play a particular role will in turn affect the self-concept.

Two methods of role learning may be delineated, i.e., intentional learning and incidental learning. Intentional learning is explicit learning and the intentional learning of a research role occurs primarily in the

university. In incidental learning, the individual adopts "the prevailing pattern." It is the author's thesis that incidental learning, much of which takes place in the clinical center, plays a significant part in the ultimate development of the self-concept of the clinician. Incidental learning occurs essentially through a process of identification (unconscious) or imitation (conscious). This places great emphasis on those in the emerging clinical psychologist's environment who have securely integrated the role of researcher into their professional self-concept. It is here that an important defect exists in the opportunities for learning a research role since the figures who securely integrate the research role into their self-concepts are frequently unavailable or minimally available in the internship. For the most part, there are figures who are diagnosticians, therapists, administrators, etc. The intern can *identify* with figures that are not in his immediate visual field, such as significant researchers in the field of psychology. However, we have perhaps failed to emphasize the fact that *imitation*, which may be more important in the development of the professional self-concept, does require the figure to be in the visual field.

This has been implicitly recognized in many clinical centers with the result that more clinical centers are today resorting to the use of research consultants who serve the important role of providing figures with a secure self-concept that incorporates the role of researcher. These are, for the most part, research psychologists from universities who visit the internship center at varying intervals, spending several hours on each visit and advising on theoretical conceptualizations, research designing, statistical analyses, and other issues pertinent to research. But even the use of consultants may be insufficient. Frequently, these are fleeting figures who appear irregularly, who may not actually become involved in research within the clinical center, and who are often not clinicians but are the prototypes of those academic figures who have originally contributed to feelings of inferiority about the clinician's research role.

²A number of the ideas presented in this paper were stimulated by Sarbin's (1954) excellent discussion of role theory.

This may be a considerable oversimplification. One may still have figures available who provide appropriate models for identification or imitation, but there may still not be a basic motivation for learning a research role. From where does the motivation to learn and practice the role of researcher stem? Such motivation may derive from factors such as financial rewards, status, advancement, and other extrinsic stimulants. However, we cannot ignore the fact that there may be more intrinsic factors operating to provide motivation for the learning of a research role. Redlich and Brody (1955) have emphasized research motivation as being derived from "powerful unconscious drives to gratify infantile curiosity and the wish for omniscience" (p. 234). Kubie (1953) has carried this further and has attempted to relate unconscious strivings to the various aspects of research such as the problem selected for study, the specific research techniques utilized, the hypothesis selected for verification, and other aspects of the research process.

In addition to learning the skills of research, including clinical research skills, and in addition to learning the role of researcher and effectively integrating it into one's self-concept, there is another type of learning that is a joint responsibility of the university and the clinical center. This is the learning of the ethics of research. I refer here to the obligations of the researcher to his subjects, to his profession, to the organization which employs him, to his collaborators, etc. Too often, one observes the researcher, be he clinical or not, whose competence in terms of knowledge and skills is obviously of a high order, but who has failed to integrate a valid system of ethical values into his research activity.

PROBLEMS OF ROLE ENACTMENT

One of the complex adjustment problems faced by the young clinician is the need to learn a multiplicity of roles, e.g., diagnostician, therapist, teacher, researchist, etc. At times, there may be conflict between certain of these roles, which may then result in the relinquishing of one of the roles in order to

resolve the conflict. Where one of the roles is that of research, this frequently leads to the relinquishing of the research role because the clinical center frequently supports this role less than others.

Role conflict occurs if the individual occupies two positions simultaneously, and when the role expectations of one position are incompatible with the role expectations of the other. It is rare indeed that situations arise which present this type of role conflict as it pertains to the practice of research. Where one finds the verbalization of such role conflict, one must suspect that there exists another explanation, other than a real role conflict, since it is rare that the individual is expected to perform a research role and any other role simultaneously. The role of clinician and the role of research worker, even where these are successfully integrated, are not performed simultaneously, but are performed at different times, in different settings, in relation to different individuals, etc. The author views the multiple roles of the clinical psychologist not as mutually incompatible, but mutually complementary. The ability to successfully perform a multiplicity of professional roles permits one to enrich the particular role that the psychologist may be enacting at a particular time.

It has frequently been stressed that the attitudes and skills required for research are significantly different from those required for clinical activity. The problem here is not that there is a basic incompatibility between research and clinical activity, but that certain types of research, usually of the experimental laboratory type, are strikingly different from what is involved in clinical practice. However, there are many problems which cannot at the moment be studied effectively in a strictly experimental laboratory manner, and here the clinical method, with its latitude, with its less standardized approach, with its greater subjectivity, can be marshaled to fulfill a research objective. The problem is one of adapting the clinical method so that it can become a useful research method. In fact, the author has stressed elsewhere the fundamental resemblance of the clinical and the scientific methods (Holzberg, 1957).

It is not being suggested that the clinical psychologist must enact all of his roles with the same degree of intensity or of involvement. It is the rare individual who can maintain intense affective involvement and intense expenditure of physical effort in all of the roles that he may be expected to perform. The clinician need not have the highest involvement in research, but somewhere in the repertoire of roles there must be integrated into the clinician the availability of the research role. Clinical psychologists vary in their enactment of the research role such that each may be located on a continuum extending from almost automatically enacting the role, to the other extreme where the clinician is very self-conscious of his own role enactment. The problem is how to make the clinician, and particularly the clinical student, become less self-conscious of the enactment of the research role.

To perform a role, the individual must clearly know the expectations of those with whom he interacts in this role. Unless the internship staff conceptualizes the intern's role as incorporating research, the intern's enactment of the research role may present many problems. Thus, he may act out a role which others may consider inappropriate, if not actually bizarre. Of crucial importance here are the overt acts of others, rather than verbalizations. If others are doing research, if research is being actively supported in the agency, this is a far more significant factor in revealing the attitudes toward a research role than mere verbalized support.

Where the professional structure demands a multiplicity of role performances as in clinical psychology, there must be "flexibility" in the individual, so that he may readily be able to shift from one role to another. It is apparent that this is one dimension of the self that may make for difficulty in the enactment of the research role, simply because the individual may be lacking in this capacity for flexibility. At the same time, the individual must possess sufficient capacity to erect barriers around a particular role being enacted, so that the role does not shift inappropriately.

DEFENSIVE REACTIONS TO RESEARCH ROLE CONFLICTS

If the actions and qualities of a given role are congruent with the self-concept maintained by the individual, this enhances the probability that the individual will be able to perform in a way consistent with the role expectations. It is the thesis of the writer that the incongruence of the role of the researchist with the self-concept of the clinician is a central problem. This is in essence a situation of conflict, and like conflicts which are not directly resolvable, leads to tension and resort to various defensive operations. One of these is rationalization, which is best exemplified by the resort to the argument that the research role is incompatible with the clinician's role.

Another defensive reaction is that of projection. This manifests itself in destructive and nihilistic attitudes toward the research of others. This may take the form of the psychologist criticizing another researcher's technical competence, his knowledge of the area, etc., all of which are projections of his own inadequacies.

Still another defense is that of denial—failing to recognize research problems or denying that problems others are posing are real research problems. This person seems to be saying: "If I do not recognize research problems, I cannot be expected to do research."

Some clinicians may resort to identification as a way of resolving the conflict. They may identify with strong clinical figures, such as psychoanalytic psychotherapists, who have considerable prestige in terms of their clinical endeavor. The clinician who can successfully make this identification may comfortably remove himself from a research role, since he can model his behavior after one who has acceptable status and prestige.

Some clinicians remain on the fringe of research. They serve as "advisers" to others but never get involved in research. Frequently, they perform a significant role as advisers, which is testimony to the fact that they have certain competencies regarding research which should lead to research activity. These individuals are so blocked by

sense of inferiority and the potential feeling that their research activity will not succeed that one finds this tragic discrepancy between genuine ability in research and an inability to function in a research role. At the other extreme is the "dilettante" who frequently does have some research interest and talent, but who wants to research every problem here and now. This frequently masks the insecurity and inferiority of the psychologist who is anxious that his research effort yield a significant contribution, and so to buttress his own security seeks to embark on a second research project simultaneously. But, even this does not assuage his anxiety so that he must turn now to a third project, hoping that in some way a multiplicity of research activities will at least yield one research project which will earn him success. This multiplicity of efforts frequently leads to activity but no sustained research.

THE CLINICIAN AS A RESEARCHER

A question that frequently is posed is whether all clinical psychologists should be expected to do research. Certainly, at the very least, we should expect that the individual who has received a PhD in clinical psychology approach his clinical tasks with an orientation geared to recognize problems that remain unanswered and questions seeking solutions. Perhaps this is as much as we should expect from the bulk of practicing clinical psychologists—this capacity to recognize and be aware of the problems needing solution. Thus, an intern may conceivably have a successful experience with regard to research if he can spend his year within a setting that, while it does not practice research, at least is self-questioning and self-analyzing with regard to its beliefs and its techniques. However, it is the thesis of the author that the more significant research experience during the internship will be accomplished where the setting is itself one where the psychologists are actively involved in research activity.

The controversy as to whether we should be training people to do research or to evaluate published research still continues. Even if we could agree that this was a legitimate conflict, in what way would the

training for doing research be different from the training that would enable one to evaluate research? Is it possible to evaluate research without systematic involvement in research that brings home to one realistically the important issues with which one is concerned in research activity?

It is important to stress that there are many skills that enter into being a well-rounded research person. Among these are: the capacity to recognize meaningful research problems; the capacity to engage in the creative thinking involved in building hypotheses and deducing implications from them; the capacity to integrate and utilize theory to refine a problem; the capacity to develop a research plan with appropriate controls that would make possible the testing of the hypotheses; the capacity to assess the nature of populations and the types of samplings to be made; the capacity to select the appropriate techniques and instruments to be used in the measuring process of the research; the capacity to deal with subjects not as clinical but as research entities; the capacity to record faithfully responses made by subjects; the capacity to tabulate, score, and analyze the data resulting from the research; and the capacity to communicate, both in written and oral form, the nature of the investigation and the results. Clearly, there are few people who possess all of these skills to a sufficient degree to qualify as the well-rounded research clinician. Here individual differences must be recognized and accepted. Individuals vary quantitatively in the extent to which they possess the various skills that have been enumerated, and a place in research exists for all people trained as clinical psychologists, each one making a contribution commensurate with his capacities and within his limitations. Group research is perhaps a prime way in permitting every clinical psychologist a research role.

RESEARCH TRAINING FOR THE INTERN

It seems imperative that the setting in which the intern receives his first major clinical experience be one in which research in all of its aspects is a vibrant part of the program. This is especially true with regard to thinking about research. Toward this end,

we have found it advisable to have frequent conferences devoted either to ongoing research or published research. Even in those settings that may not be able to integrate internes into ongoing research, it seems advisable that this provision for regular discussions should become institutionalized as part of the program.

One problem that presents itself in many clinical settings is that of providing time for research activity in the face of service demands. We have utilized the technique of releasing the intern every third week from diagnostic responsibilities in order to permit him to work on research. The importance of providing time is that it is a concrete demonstration of the department's and the agency's attitudes toward research. This open acceptance of research activity as a legitimate function of the clinician will minimize the anxiety or guilt in the individual who feels that he is sacrificing services for patients by engaging in research.

Another problem with which we have struggled has been that of organizing interns so that they may participate in group research which provides a richer learning experience for the intern than he can normally expect from individual research. The opportunity for continued conferences on the group research introduces a significant dimension into the training of the intern that can hardly be matched when the intern is working alone on his own problem, even where he is being supervised. Working in the group on a group research does not minimize the opportunity for individual supervision or for contact with consultants, but rather adds to this the involvement with a group of other psychologists with whom the intern can interact on research in a continuous way over a year. Besides the potentially greater intellectual stimulation made possible by group research, there is the support that is provided in dealing with the many complex problems that arise in the course of research activity.

The use of research consultants is felt to be a very valuable part of the research training experience for the intern. Not only do they add to the research atmosphere of the department, but also these people, when carefully

selected, can provide genuine help on research problems from the theoretical conceptualization to the technical aspects of designing the research and analyzing the data. The use of these consultants may also help the intern to understand that there may not be as great a cleavage between the university research psychologist and the "field" clinician as many interns seem to feel. By observing the fact that these two can communicate, share, and sometimes collaborate to their mutual benefit, the intern learns that there need not be this sharp schism between clinical activity and research as it is often defined in the university research center.

A problem that must frequently be faced is the latent feeling in some interns that, if they engage in agency research, they will somehow be misused. This misuse may occur at two levels: (a) The intern may fear that he will be relegated to tasks that are either demeaning or more appropriate to an individual with less training. This suggests that, in providing a training experience in research for the intern, it must be an experience that will be truly an educational one. He cannot be used simply as an assistant or as a clerical worker, but must share directly in the research from its conceptualization to its execution. (b) The intern may fear that he will not be rewarded for his contributions to the research in the form of authorship. In clinical activity, this concern is not present. Whether he is performing as a diagnostician or therapist, the intern is explicitly recognized for his services. Since the diagnostic workup and the record of treatment of the patient in the medical file bear his name, this serves to provide the intern with formal recognition of his labors. In research, this immediate recognition is not readily apparent. It has been our experience that with some interns this is a significant concern that may contribute to the intern's reluctance to become involved in hospital research. We have found that it is well to be explicit with interns with regard to the rewards that they may expect from participation in the research. This has most usually meant that the intern is accepted as a collaborator from the point of view of authorship.

With the provision for the intern to work on a clinical research problem during the in-

ternship, there carries with it a requirement that such research should be supervised by someone who not only can assess the research from a technical point of view (design, sampling, etc.), but can also make the research as clinically meaningful as possible. It is the rare intern who can function independently with regard to research activity, even when he has already completed his doctoral dissertation. We have recently introduced, on a trial basis, a research training program for interns which provides for the joint supervision of interns' research by a university experimentalist and by a clinician.

Part of the responsibility of the research worker is to communicate his research and its findings effectively both in written and in oral form. It is part of the training of the intern that once he is involved in his research he is expected to prepare reports to be presented at group conferences at which consultants may be present.

CONCLUSIONS

It may be possible to reconceptualize the problem of how one trains the intern to perform a research role effectively. In a sense, we may say that the individual learns to do research most effectively when he is globally involved in this process of learning. That is, he learns first by feeling that research is an activity personally and professionally reward-

ing and by feeling comfortable in performing this research role because it does not conflict with other roles or with his own self-concept; he learns by thinking research, by being involved in activities that will permit him to extend his intellectual horizons with regard to research and particularly clinical research; and finally, he learns by doing research, by actually becoming involved in research activity. Our emphasis here has been that the internship's responsibility with regard to developing the research role of the clinician must be to permit the intern to begin to *feel* like a research person, to provide experiences that will extend his *thinking* about clinical problems and methods for dealing with them, and to provide an experience of *doing* in which he may practice the skills of a research worker.

REFERENCES

- HOLZBERG, J. D. The clinical and scientific methods: Synthesis or antithesis? *J. proj. Tech.*, 1957, 21, 227-242.
- KUBIE, L. S. Some unresolved problems of the scientific career. *Amer. Scientist*, 1953, 41, 596-613.
- REDLICH, F. C., & BRODY, E. B. Emotional problems in interdisciplinary research in psychiatry. *Psychiatry*, 1955, 18, 233-239.
- SARBIN, T. R. Role theory. In G. Lindzey (Ed.), *Handbook of social psychology*. Cambridge: Addison-Wesley, 1954.

(Received April 11, 1960)

COMPARABILITY OF INTELLIGENCE QUOTIENTS OF MENTAL DEFECTIVES ON THE WECHSLER ADULT INTELLIGENCE SCALE AND THE 1960 REVISION OF THE STANFORD-BINET¹

GARY M. FISHER,² BEVERLY A. KILMAN, AND ANNA M. SHOTWELL

Pacific State Hospital, Pomona, California

A number of studies comparing the Wechsler-Bellevue (W-B) and the Wechsler Intelligence Scale for Children (WISC) with the Stanford-Binet (S-B) have been reported. Correlation coefficients between IQs from these Wechsler scales and the S-B have been of the same order (.6 to .9) in normal, neuropsychiatric, and mentally retarded populations (Alderdice & Butler, 1952; Benton, Weider, & Blauvelt, 1941; Goldfarb, 1944; Gothberg, 1949; Mitchell, 1942; Nale, 1951; Sandercock & Butler, 1952; Stacey & Levin, 1951; Wechsler, 1944). Although some studies have indicated IQs from the WISC to be slightly higher than those from the S-B, the two instruments give fairly comparable scores (Frandsen & Higginson, 1951; Gehman & Matyas, 1956; Harlow, Price, Tatham, & Davidson, 1957; Weider, Noller, & Scramm, 1951). A number of studies have indicated that W-B IQs are consistently higher than S-B IQs of older subjects (Halpern, 1942; Mitchell, 1942) and that this discrepancy becomes larger as age increases and intellectual level decreases (Bensberg & Sloan, 1950; Kutash, 1945; Mundy & Maxwell, 1958).

Only one study, that of Wechsler (1958),

compares the Wechsler Adult Intelligence Scale (WAIS) with the S-B. In a sample of 52 male prisoners, aged 16 to 26 years, the mean S-B IQ was five points higher than the mean WAIS IQ, and the correlation between the two IQs was .85. The present study compares adult hospitalized retardates of varying age and intelligence level as to their IQs on the WAIS and the 1960 revision of the Stanford-Binet Form L-M.

METHOD

The sample consisted of 180 mentally retarded subjects in three California state hospitals who were 18 years or older and who had a diagnosis of familial or undifferentiated mental retardation.³ The sample was classified into four unequal age groups and three unequal IQ categories, in order to obtain 12 equal-sized subgroups of 15 patients each. The age grouping by years was: 18-34, 35-44, 45-54, and 55-73. The grouping by intellectual level (based on the average of WAIS and S-B IQ)⁴ was: 46 and below, 47-54, and 55 and over.

Half of the subjects were given the WAIS first and the remainder the S-B first. For a given subject, different examiners administered the two tests. A counterbalancing procedure was employed in assigning subjects to the two testers in order to control for possible examiner difference. To control for changes

¹ This study was supported by a grant from the California Department of Mental Hygiene (No. R-58-14.1-3). The authors gratefully acknowledge the assistance of Edward Bartsch, Milton Dooley, James Judge, Robert Kirby, Earl Owens, Dorian Rose, and Charles Windle. Special thanks are due Maud Merrill for supplying material for the 1960 Revision of the Stanford-Binet before its publication (Terman & Merrill, 1960) and to Arthur Silverstein for his assistance with the statistical analyses.

² Now at Fairview State Hospital, Costa Mesa, California.

³ Thus subjects with known brain damage were excluded from the study.

On the WAIS, a subject must achieve a minimal Total Scaled score of 11 to obtain an IQ from the table of norms. Consequently, any subject achieving less than this minimum score was excluded from the sample since only prorated IQs could be estimated and presumably the WAIS would not be an appropriate test for such a subject.

⁴ This procedure for determining IQ level was adopted in order to circumvent the problem of statistical regression toward the mean which would occur if IQ level were determined by only one test.

TABLE 1
COMPARISON OF WECHSLER ADULT INTELLIGENCE SCALE AND
STANFORD-BINET IQs FOR EACH AGE LEVEL

Age	WAIS IQ			S-B IQ			<i>t</i> value	Correlation WAIS & S-B
	Mean	SD	Range	Mean	SD	Range		
18-34	58.42	8.40	45-77	44.02	9.25	26-68	15.00	.736
35-44	59.44	10.20	45-96	45.84	10.99	31-75	12.04	.752
45-54	60.29	8.31	48-76	43.51	8.61	26-68	18.64	.752
55-73	63.56	9.88	51-87	41.00	8.33	26-57	23.92	.777

which might occur in intellectual functioning over an extended period of time, all subjects were given both tests within a 12-month period, with the great majority of subjects receiving both tests within a 4-month period.

To secure data that might help in developing an hypothesis regarding the relation of social competency to intellectual functioning on the WAIS and S-B, all subjects were administered the Vineland Scale of Social Maturity (Doll, 1953). In addition, an adapted form of a Scale of Minimal Social Behavior (Farina, Arenberg, & Guskin, 1957) was administered, but because of the lack of variability in scores from this instrument no use could be made of the results.

RESULTS

An analysis of variance of the difference scores between WAIS and S-B IQs was performed. The results indicated that age was significant in determining the magnitude of the discrepancy between the two IQs ($F = 16.74$; $df = 3$; $p < .001$), but neither IQ level nor the interaction between IQ level and age was significant.⁵

Table 1 shows the comparison of WAIS and S-B IQs for each age level. Of the 180 subjects examined only 3 had an S-B IQ higher than their WAIS IQ, the largest difference being five points. The *t* values for the significance of the difference between correlated means indicated that the mean WAIS IQ was significantly larger (beyond the .001 level) than the S-B IQ at each age level. The correlation coefficients between WAIS and

S-B IQs ranged from .736 to .777 and a chi square test (Edwards, 1950) indicated that the differences among them were not significant ($\chi^2 = .20$; $df = 3$; $p > .95$).

Table 2 sets forth the difference scores between WAIS IQ and S-B IQ by age level. The *t* tests indicated that the mean difference scores for the three younger age groups were significantly smaller ($< .01$) than for the oldest age group. Between the ages of 18 and 54 years, the WAIS IQ averaged 15 points higher than the S-B IQ, whereas in subjects over 55 years the difference averaged 23 points. The standard deviation of the difference score was approximately 6. The WAIS, unlike the S-B, takes systematic account of the lower performance of older adults; hence, the older the adults (beyond a peak between 25 and 30) the lower the absolute level of performance and hence the larger the discrepancy between S-B and WAIS. As the WAIS standardization population included adults of a wide age range whereas the S-B did not, it would be logical to assume that the WAIS IQ is a more accurate measure of in-

TABLE 2
DIFFERENCE SCORES BETWEEN WECHSLER ADULT
INTELLIGENCE SCALE IQs AND STANDARD-BINET IQs

Age	Difference Scores		
	Mean	SD	Range
18-34	14.49	6.49	-5 to 32
35-44	13.60	7.53	-5 to 27
45-54	16.56	5.75	5 to 32
55-73	22.56	6.28	10 to 36

Note.—For each subject, the S-B IQ was subtracted from the WAIS IQ.

⁵ Analyses of the difference scores between WAIS Verbal IQ and S-B IQ, and between WAIS Performance IQ and S-B IQ were also performed. The results were virtually identical with the analysis of the difference scores between WAIS and S-B IQs. Moreover, these two sets of difference scores were of the same magnitude as difference scores between WAIS Full Scale IQs and S-B IQs.

TABLE 3

COMPARISON OF IQs FROM THE WECHSLER ADULT INTELLIGENCE SCALE AND STANFORD-BINET WITH SOCIAL AGES FROM THE VINELAND SCALE OF SOCIAL MATURITY

Age	WAIS Mean IQ	S-B Mean IQ	Vineland Scale of Social Maturity			Correlation Coefficients	
			Mean Age	SD	Range	WAIS IQ & Social Age	S-B IQ & Social Age
18-34	58.42	44.02	10-2	1-6	5-6 to 13-3	.265	.322
35-44	59.44	45.84	10-2	1-9	5-5 to 13-0	.438	.171
45-54	60.29	43.51	9-9	2-0	4-0 to 12-10	.638	.433
55-73	63.56	41.00	9-7	2-1	5-5 to 13-3	.534	.670

telligence. It has been suggested, however, that there is a spurious allowance for "normal" deterioration with age for retardates on the Wechsler scales (Bensberg & Sloan, 1950). Until validity studies, preferably of the longitudinal type, of the WAIS with retarded populations are made, the problem of differential rate of intellectual decline depending on level of intellectual functioning remains unanswered.

The regression equations for predicting the IQ on one test from that on the other for the three younger age groups appeared sufficiently similar to warrant the use of the following regression equations for subjects between 18 and 54 years:

Predicting WAIS IQ from S-B IQ:

$$Y = .68X + 29.15$$

$$(SE_e = 6.15)$$

Predicting S-B IQ from WAIS IQ:

$$Y = .79X - 2.45$$

$$(SE_e = 6.61)$$

For subjects 55 years and older, the following regression equation should be used:

Predicting WAIS IQ from S-B IQ:

$$Y = .92X + 25.84$$

$$(SE_e = 6.22)$$

Predicting S-B IQ from WAIS IQ:

$$Y = .65X - .31$$

$$(SE_e = 5.25)$$

In order to gain some knowledge of the relation between social competency and the IQs from the two intelligence scales, comparisons of the Social Ages (SA) from the Vineland Scale of Social Maturity were made with the IQs from the WAIS and the S-B. These data are presented in Table 3. Chi square tests indicated that the correlation coefficients between WAIS IQ and Vineland SA and between S-B IQ and Vineland SA were homogeneous for the four age levels ($\chi^2 = 5.249$ and 9.264 , respectively; $df = 3$; $p > .01$). For the total sample the correlation coefficient between Vineland SA and WAIS IQ was .450, and between Vineland SA and S-B IQ was .405. A nonsignificant t value of .939 for the significance of the difference between correlated correlations indicated that the WAIS and S-B correlated equally well with Vineland SA. The data shown in Table 3 suggest a trend for WAIS IQs to increase, and S-B IQs and Vineland SA to decrease, with age. According to analyses of variance these observed trends all proved to be nonsignificant at the .01 level (WAIS: $F = 2.54$; S-B: $F = 2.01$; Vineland SA: $F = 1.29$). More thorough investigation would be necessary in order to answer the question of the relation between social competency and intelligence as measured by these two scales.

SUMMARY

This study sought to determine the effect of age and level of retardation on the comparability of IQs from the Wechsler Adult Intelligence Scale and the 1960 revision of the

Stanford-Binet. In addition, a measure of social competency was related to the IQs from the two scales. It was determined that age, but not level of retardation, was significant in determining the magnitude of the difference between WAIS and S-B IQs. WAIS IQs averaged 15 and 23 points higher than S-B IQs for subjects 18-54 years and 55-73 years, respectively. Regression equations were calculated to translate the IQ from one test to the other test. The WAIS and S-B IQs correlated equally well with Social Ages from the Vineland Scale of Social Maturity.

REFERENCES

- ALDERDICE, E. T., & BUTLER, A. J. An analysis of the performance of mental defectives on the Revised Stanford-Binet and the Wechsler-Bellevue. *Amer. J. ment. Defic.*, 1952, 56, 609-614.
- BENSBERG, G. J., & SLOAN, W. Wechsler's concept of "normal deterioration" in older mental defectives. *J. clin. Psychol.*, 1950, 6, 359-362.
- BENTON, A. L., WEIDER, A., & BLAUVELT, J. Performance of adult patients on the Bellevue intelligence scales and the revised Stanford-Binet. *Psychiat. Quart.*, 1941, 15, 802-806.
- DOLL, E. A. *Measurement of social competence*. Minneapolis: Educational Publishers, 1953.
- EDWARDS, A. L. *Experimental design in psychological research*. New York: Rinehart, 1950.
- FARINA, A., ARENBERG, D., & GUSKIN, S. A scale for measuring minimal social behavior. *J. consult. Psychol.*, 1957, 21, 265-268.
- FRANDSEN, A. N., & HIGGINSON, J. B. The Stanford-Binet and Wechsler Intelligence Scale for Children. *J. consult. Psychol.*, 1951, 15, 236-238.
- GEHMAN, I. A. H., & MATYAS, R. P. Stability of the WISC and Binet tests. *J. consult. Psychol.*, 1956, 20, 150-152.
- GOLDTARB, W. Adolescent performance in the Wechsler-Bellevue and the revised Stanford-Binet examination, Form L. *J. educ. Psychol.*, 1944, 35, 503-507.
- GOTHBERG, LURA C. A comparative study of the Stanford-Binet old form test and the Wechsler-Bellevue verbal, performance, and full scale as shown in the results of unselected employees. *Amer. J. ment. Defic.*, 1949, 53, 497-503.
- HALPERN, FLORENCE. Comparison of the revised Stanford-Binet with the Bellevue adult intelligence scale. *Psychiat. Quart. Suppl.*, 1942, 16, 206-211.
- HARLOW, J. E., PRICE, A. C., TATHAM, L. J., & DAVIDSON, J. F. Preliminary study of comparison between Wechsler Intelligence Scale for Children and Form L of revised Stanford-Binet scale at three age levels. *J. clin. Psychol.*, 1957, 13, 72-73.
- KUTASH, S. B. A comparison of the Wechsler-Bellevue and the revised Stanford-Binet scales for adult defective delinquents. *Psychiat. Quart. Suppl.*, 1945, 19, 677-685.
- MITCHELL, M. B. Performance of mental hospital patients on the Wechsler-Bellevue and revised Stanford-Binet Form L. *J. educ. Psychol.*, 1942, 33, 538-544.
- MUNDY, LYDIA, & MAXWELL, A. E. The assessment of the feeble-minded. *Brit. J. med. Psychol.*, 1958, 31, 201-211.
- NALE, S. The children's Wechsler and the Binet on 104 mental defectives at the Polk State School. *Amer. J. ment. Defic.*, 1951, 56, 419-423.
- SANDERCOCK, MARIAN G., & BUTLER, A. J. An analysis of the performance of mental defectives on the Wechsler Intelligence Scale for Children. *Amer. J. ment. Defic.*, 1952, 57, 100-105.
- STACEY, C. L., & LEVIN, JANICE. Correlation analysis of scores of subnormal subjects on the Stanford-Binet and Wechsler Intelligence Scale for Children. *Amer. J. ment. Defic.*, 1951, 55, 590-597.
- TERMAN, L. M., & MERRILL, MAUD A. *Stanford-Binet intelligence scale*. Boston: Houghton Mifflin, 1960.
- WECHSLER, D. *The measurement of adult intelligence*. (3rd ed.) Baltimore: Williams & Wilkins, 1944.
- WECHSLER, D. *The measurement and appraisal of adult intelligence*. (4th ed.) Baltimore: Williams & Wilkins, 1958.
- WEIDER, A., NOLLER, P. A., & SCRAMM, T. A. The WISC and the Stanford-Binet. *J. consult. Psychol.*, 1951, 15, 330-333.

(Received April 1, 1960)

PERCEPTUAL SIZE CONSTANCY IN CHRONIC SCHIZOPHRENIA¹

H. W. LEIBOWITZ

University of Wisconsin

AND VLADIMIR PISHKIN

Veterans Administration Hospital, Tomah, Wisconsin

The ability to judge correctly the sizes of objects despite variation in viewing distance, i.e., size constancy, represents an important biological achievement of living organisms and has been studied in relation to the underlying mechanisms as well as to psychopathology. In particular, the possibility that size constancy differs among schizophrenics as compared with normals has been of interest to a number of investigators and theorists (Lovinger, 1956; Maes, 1957; Raush, 1952, 1956; Reynolds, 1954; Sanders & Pacht, 1952; Weckowicz, 1957). Bruner (1951) states the theoretical problem most succinctly when he suggests that a withdrawal from object relations and an increasing concern for the self might lead to a breakdown in perceptual constancy. However, as pointed out in a recent summary (Rabin & King, 1958), there is little agreement among the results of previous experiments. Sanders and Pacht (1952) found "overconstancy" among a schizophrenic group as compared with normals, while Weckowicz (1957) reports the opposite relationship. Raush (1956) found no differences between nonparanoid schizophrenics, although his paranoid group did differ from normals.

The purpose of the present experiment is to re-examine this problem, utilizing a group of chronic, undifferentiated schizophrenics chosen from the patient population of a neuropsychiatric hospital to be most "withdrawn" with respect to their behavior. It is felt that

the selection of the extreme cases from among a group of chronic patients whose diagnosis is more reliable provides an excellent test of the hypothesis under consideration. The method of testing has been modified after techniques which have been employed in experimental laboratories and for which a large body of normative data is already available (Holway & Boring, 1941; Leibowitz, in press; Leibowitz, Chinetti, & Sidowski, 1956). Essentially, the technique requires the subject to signal which of two sticks is larger, a simple task which has been used successfully with children and feeble-minded groups. Data are obtained over a wide range of viewing distances, thus providing the basis for a more complete analysis of the size matching-distance relationship.

METHOD

Subjects

The members of the experimental group were 35 male patients of the Veterans Administration Hospital in Tomah, Wisconsin, carrying a diagnosis of chronic, undifferentiated schizophrenia. There was no history of brain damage, organic pathology, or visual defect. None of the patients were receiving EST, although many of them were on different types of drug therapy. They were selected by the psychiatric ward team to be mainly characterized by withdrawal symptoms. The mean age of the 35 patients was 39.41 years with a range from 24 years to 56 years. The average length of psychiatric hospitalization was 8.84 years with a range of from 2 years to 13 years. The members of the control group were 20 psychiatric aides and were selected at random. Their mean age was 38.52 years with a range of from 25 years to 54 years.

Experimental Conditions

The experiment was conducted in a hospital corridor 9½ feet wide, 9½ feet high, and 135 feet in

¹ Supported, in part, by Grant M-1090 from the National Institute of Mental Health, National Institutes of Health, United States Public Health Service, and the Veterans Administration Research Program.

length. Large windows were located every 12 feet in both walls of the corridor and all tests were made in daylight. The test objects consisted of five wooden dowels cut from 1 inch stock to lengths of 2, 4, 8, 16, and 24 inches. These were mounted in square bases and 4 inches by 4 inches and were painted "flat" black. The viewing distance of the test objects was proportional to their size, the ranges being 10, 20, 40, 80, and 120 feet, respectively. Thus, the visual angle subtended by each object, and the size of the corresponding retinal image, was the same in each case (0.96 degree of arc). The subject was seated at one end of the corridor and first shown one of the test objects set up at its appropriate viewing distance. The subject was told to look at the height of the stick and to tell whether a comparison stick, which was placed at right angles to the line of sight to the test object and 8 feet from the subject, was shorter or taller than the test object. The comparison objects were chosen from a series ranging in size from 1 inch to 35 inches and were indistinguishable, except for size, from the test objects. The order of presentation of the comparison stimuli was chosen so that the first comparison would easily elicit a "higher" or "shorter than" response from the subject, while the second reversed his response. Subsequent sticks were chosen to be progressively nearer the point of subjective equality. The order of presentation of the comparison objects was determined by a 5×5 latin square design. No difficulty was encountered in testing the patients. They cooperated well, exhibited no difficulty in understanding the instructions nor in making judgments, and showed no signs of complaint throughout the study, which lasted approximately 15 minutes for each patient. The two authors served alternately as experimenters.

RESULTS

The mean sizes of the comparison objects which matched the test objects at the various distances employed are given for both groups along with their standard deviations in Table 1. Figure 1 represents a plot of mean matched size as a function of test object distance. On this figure a horizontal line represents a prediction in terms of the "law of visual angle,"

TABLE 1
MATCHED SIZE OF A SERIES OF TEST OBJECTS AS A FUNCTION OF DISTANCE FOR SCHIZOPHRENIC AND NORMAL SUBJECTS

Test Object Size (Inches)	Test Object Distance (Feet)	Schizophrenic ($N=35$)		Normal ($N=20$)	
		Mean	SD	Mean	SD
2	10	2.08	0.23	2.13	0.22
4	20	4.11	0.19	4.11	0.36
8	40	9.09	1.09	8.97	1.08
16	80	17.87	1.74	17.03	1.88
24	120	24.71	2.99	25.48	2.41

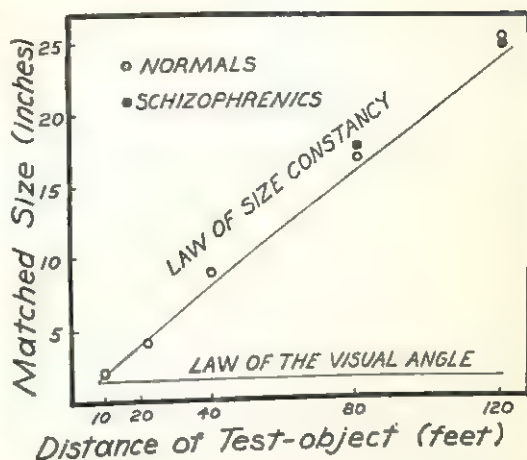


FIG. 1. Matched size as a function of distance for a group of normals and of chronic schizophrenics. (The test objects subtended the same visual angle at all distances of observation.)

a condition which would obtain if judged size depended only on retinal image size. In the present study, retinal image size was constant so that matched size, according to this prediction, would be the same for all test objects and viewing distances. The other theoretical extreme, the "law of size constancy," is represented by the diagonal line. According to this prediction, perceived size is independent of distance so that matched size would be equal to the actual size of the test objects. As is typical in such experiments conducted with binocular vision, adequate illuminance, and a well articulated visual field, the data for normals lie very close to the prediction in terms of size constancy. The data for the schizophrenics also lie close to this theoretical line.

The nature of the data indicated the use of a nonparametric test. The Mann-Whitney U test was applied (Siegel, 1956). From this analysis it was concluded that size judgments of the schizophrenic group do not differ significantly from the normal sample on this task ($U = 119$, $p < .424$). In addition, respective variabilities of the two groups were compared. As expected, none of the differences reached significance.

DISCUSSION

The results of the present study indicate no difference between the ability of chronic schizophrenics and normals to judge correctly

the sizes of test objects from 10 to 120 feet. Both groups exhibit a high degree of size constancy. These data are in agreement with those of Raush (1952), who found essentially the same results with nonparanoid schizophrenics. The disagreement of the present data, as well as those of Raush, with other investigators could well be a result of differences in the patient population (e.g., Sanders & Pacht, 1952, tested outpatients) or in the procedure employed. In any event, it is important to recognize that even in those studies which do differentiate between normals and schizophrenics the magnitude of the differences are small in relation to differences found as a result of variation of less subtle variables, such as the instructions given the subjects (Gilinsky, 1955; Holaday, 1933), or the age of the subjects (Beryl, 1926; Zeigler & Leibowitz, 1957).

The present data have implications in relation to psychopathology, as well as to a theoretical understanding of size constancy. With respect to the chronic schizophrenic group tested, it is clear that whatever disorders of perception, thinking, or behavior they suffer from, their size constancy is unaffected. Indeed, it would be difficult to imagine such patients, who move about the hospital grounds and engage in sports, judging incorrectly the sizes of environmental objects. The difficulty with schizophrenics, as seen clinically, may not be in relation to neutral objects, such as the sticks utilized in this study, but to other individuals and to their own patterns of thinking and emotions. Apparently, the withdrawal characteristic of schizophrenia, in general, and of the subjects utilized in this study, in particular, may be significant in interpersonal relationships only, and are not indiscriminately employed as defenses in purely nonaffective areas, such as size judgment. As suggested by Bruner (1951), one would be more likely to find differences between schizophrenics and normals if the "test objects" were human beings.

The present size matching data are also of importance in relation to the mechanisms which subserve the size constancy effect. It has been demonstrated that the size con-

stancy of children for near objects is good, but that the ability to judge correctly the sizes of distant objects develops slowly (Zeigler & Leibowitz, 1957). These data suggest that there are a number of separate mechanisms involved in size constancy, the nearer distances being mediated by the kinesthetic cues of accommodation and convergence, the farther by a more slowly developing perceptual learning process (Leibowitz & Moore, 1960). The available data indicate that this learning process is complete by the early teens or, in any event, since the patients in the present study were veterans, before the age at which the subjects tested were hospitalized. The lack of any difference between the groups in the present experiment suggests that the mechanisms underlying size constancy for neutral objects, such as used in this study, are independent of personality changes even as severe as those encountered in schizophrenia. This conclusion is in agreement with similar studies which demonstrate that the development of size constancy is independent of intellectual processes as indicated by the lack of differences between feeble-minded and normal subjects on size matching tasks (Jenkin & Morse, in press; Leibowitz, in press). Thus, it would appear that the development of size constancy, although closely related to the age of the subject, is independent of mental and personality development.

SUMMARY

The ability to judge object size as a function of distance was determined for a group composed of chronic, undifferentiated schizophrenics, as well as a control group of psychiatric aides.

There were no significant differences in the matches produced by the two groups. Both judged correctly the sizes of the test objects at all distances.

It is suggested that the absence of any differences is due to the fact that size matching requires abilities which are fully developed prior to the onset of schizophrenia and which are unaffected by the characteristic withdrawal observed in this pathology.

REFERENCES

- BERYL, F. Über die Grössenauffassung bei Kindern. *Z. Psychol.*, 1926, 100, 344-371.
- BRUNER, J. S. Personality dynamics and perceiving. In R. Blake & G. Ramsey (Eds.), *Perception: An approach to personality*. New York: Ronald, 1951. Pp. 121-145.
- GILINSKY, A. The effect of attitude upon the perception of size. *Amer. J. Psychol.*, 1955, 68, 173-192.
- HOLADAY, B. E. Die Grössenkonstanz der Schdinge bei Variation der inneren und äusseren Wahrnehmungsbedingungen. *Arch. ges. Psychol.*, 1933, 88, 419-486.
- HOLWAY, A. H., & BORING, E. G. Determinants of apparent visual size with distance variant. *Amer. J. Psychol.*, 1941, 54, 21-37.
- JENKIN, N., & MORSE, S. Developmental and intellectual processes in size-distance judgment. *Amer. J. Psychol.*, in press.
- LEIBOWITZ, H. Apparent visual size as a function of distance for mentally deficient subjects. *Amer. J. Psychol.*, in press.
- LEIBOWITZ, H., CHINETTI, P., & SMOWSKI, J. Exposure duration as a variable in perceptual constancy. *Science*, 1956, 123, 668-669.
- LEIBOWITZ, H., & MOORE, D. The role of oculomotor adjustments in the perception of size. *J. Opt. Soc. Amer.*, 1960, 50, 507. (Abstract)
- LOVINGER, E. Perceptual contact with reality in schizophrenia. *J. abnorm. soc. Psychol.*, 1956, 52, 87-91.
- MAES, J. L. Size constancy in schizophrenia. Unpublished master's thesis, Michigan State University, 1957.
- RABIN, A. I., & KING, G. F. Psychological studies. In L. Bellak (Ed.), *Schizophrenia*. New York: Logan, 1958.
- RAUSH, H. L. Perceptual constancy in schizophrenia: I. Size constancy. *J. Pers.*, 1952, 21, 176-187.
- RAUSH, H. L. Object constancy in schizophrenia: The enhancement of symbolic objects and conceptual stability. *J. abnorm. soc. Psychol.*, 1956, 52, 231-234.
- REYNOLDS, G. A. Perceptual constancy in schizophrenics and "normals." Unpublished doctoral dissertation, Purdue University, 1954.
- SANDERS, R., & PACT, A. R. Perceptual size constancy of known clinical groups. *J. consult. Psychol.*, 1952, 16, 440-444.
- SIEGEL, S. *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill, 1956.
- WECKOWICZ, T. E. Size constancy in schizophrenic patients. *J. ment. Sci.*, 1957, 103, 475-486.
- ZEIGLER, H. P., & LEIBOWITZ, H. Apparent visual size as a function of distance for children and adults. *Amer. J. Psychol.*, 1957, 70, 106-109.

(Received April 11, 1960)

THE RELATIONSHIPS BETWEEN INDIVIDUALLY DEFINED AND GROUP DEFINED SOCIAL DESIRABILITY AND PERFORMANCE ON THE EDWARDS PERSONAL PREFERENCE SCHEDULE

ALFRED B. HEILBRUN, JR. AND LEONARD D. GOODSTEIN

State University of Iowa

Since its publication in 1954, the Edwards Personal Preference Schedule (EPPS) has been the focus of considerable research bearing upon the problem of social desirability as a source of response variance. The EPPS was constructed so that each item consists of a pair of statements matched for social desirability, thus presumably forcing the test taker to respond in accord with his actual behavioral characteristics rather than the social desirability of the statements. To the extent that Edwards (1959b) is correct in considering social desirability as contributing principally to measurement error, studies of the relationship between social desirability and EPPS performance are important in an overall assessment of the test's validity. However, the many investigations have provided rather inconsistent results and conclusions.

Several studies have concluded that social desirability has been eliminated as an important determinant of performance on the EPPS. Navran and Stauffacher (1954) found a near-zero correlation between social desirability ratings of the traits being measured and scores on the EPPS scales. Borislow (1958) reported that subjects were able to change their test profiles under both social and personal desirability faking sets but concluded that desirability had been eliminated as a source of conscious response manipulation on the EPPS because no consistent scale patterns among the subjects were obtained under either set. Finding only a "slightly greater than chance" number of significant point biserial correlations between the subjects' Social Desirability scale scores on the

MMPI and frequency of choice of the first or second statement in each item pair of the EPPS led Kelleher (1958) to conclude that social desirability plays an insignificant role in EPPS performance.

In contrast to these negative findings, several investigators have reported results which indicate that desirability remains an important source of response variance on the EPPS. Heilbrun (1958) found a .60 correlation between EPPS scale scores and *personal* desirability of the test scales. Corah, Feldman, Cohen, Gruen, Meadow, and Ringwall (1958) obtained a .88 correlation between the percentage of college subjects judging the first statement in an item as more socially desirable and the percentage of similar subjects in an independent group who endorsed that statement as self-characteristic. Edwards, Wright, and Lunneborg (1959), pointing out that Corah et al. used only 30 EPPS items taken out of context, have repeated the procedure using all items in the test. They report somewhat lower correlations (.69 and .61) for two samples, but the relationship between desirability and endorsement still remained.

One common characteristic of all of these studies described above is that they have sought to evaluate the relationship between desirability and EPPS performance using some type of *group* desirability statistic (e.g., group average or group percentage) in the analysis. It is possible that if the relationship between each subject's *individual* desirability ratings of the statement alternatives in the EPPS and his actual statement selections were determined, a more accurate inference could

be made regarding the influence of social desirability. The logical expectation would be that if social desirability is related to EPPS performance, this should be more evident using an individualized approach, since the individual desirability values which each subject assigns to test statements should be more relevant in predicting *his* statement endorsements than would averaged group values. A recent study by Taylor (1959), however, found the opposite to be the case. He determined the correlations between both individual and group social desirability values and endorsement of MMPI items for schizophrenic patients and found the group-averaged values correlated considerably higher ($r = .79$) with performance than did the individual values ($r = .36$).

The present study had three purposes: (a) to relate individual social desirability values and EPPS performance to clarify the extent to which individual social desirability set contributes to response selection; (b) to compare the prediction of EPPS responses from individual social desirability values with prediction of these responses from group social desirability values; and (c) if individual values are found to be related to overall EPPS performance, to evaluate the specific relationship between individual desirability and each EPPS scale.

METHOD

Subjects. The 58 subjects used in this study were obtained from a large undergraduate psychology course at the State University of Iowa. This sample included 29 males and 29 females.

Test. The EPPS (Edwards, 1959a) is an objective, rationally derived, multivariate, personality questionnaire. It includes 135 different statements bearing upon personality characteristics, 9 for each of the 15 traits measured by the test. These statements are arranged in 210 pairs with each pair including 2 statements matched for social desirability but measuring different traits. The final score for each trait represents the number of occasions out of a possible 28 on which the subject has selected that trait statement as more characteristic than the paired statement.

Procedure. All subjects were initially group administered the EPPS under standard instructions. About 2 months later the same subjects were given the 135 different statements included in the test and asked to rate each as a trait in other people on a nine-point scale from highly socially undesirable to highly socially desirable. This rating procedure was

identical to that described by Edwards (1957) in his initial derivation of statement desirability values. Having obtained the subject's EPPS responses under standard conditions and individual desirability ratings from the nine-point scale for both statements in each item pair, it was possible to determine the percentage of times the subject endorsed as self-characteristic the statement having the higher individual social desirability value *for him*. Items for which the subject had ascribed equal individual social desirability values to both statements were omitted from the analysis.

The second step was to determine the percentage of times each subject endorsed as self-characteristic the more highly socially desirable statement in each item pair where the desirability values were those assigned to the statements by Edwards in his initial construction of the test. This percentage reflects the correspondence between EPPS response and group social desirability values obtained from an independent group of college students. There are 204 of the 210 EPPS items for which the group values of the paired statements differ.

Finally, the values assigned to the nine statements measuring each of the 15 EPPS variables were averaged for each subject, and these mean values defined the subject's individual social desirability ratings for the 15 traits. High and low social desirability groups were then constituted independently for each trait and the mean scale scores for these high and low groups were compared to assess whether the relationship between individual desirability and EPPS performance varied over scales. Since there are established sex differences on the EPPS scales (Edwards, 1959a), precautions were taken to insure approximately equal numbers of males and females in both desirability groups for each scale. This was accomplished by distributing individual social desirability scores for each scale separately by sex and defining high and low groups by cutting at or near the median for both sex distributions. The two high and two low desirability groups for each scale were then recombined into one high and one low group with nearly equal numbers of males and females.

RESULTS

Individual social desirability values and EPPS performance. There was a mean of 150.97 items out of a possible 210 in which the paired statements had discrepant individual social desirability values for the 58 subjects in this study. For those items within which the statement values varied, the subjects averaged endorsing the individually more desirable statement as more self-characteristic 67.16% of the time. This percentage value differs significantly from a 50% chance expectancy ($t = 10.34$ for 57 *df*, $p < .001$) and clearly indicates that performance on the EPPS is related to the individual ratings of

TABLE 1

COMPARISON BETWEEN EPPS SCALE SCORES FOR THE 15 SEPARATE HIGH AND LOW SOCIAL DESIRABILITY GROUPS

EPPS Scale	High Social Desirability			Low Social Desirability			<i>t</i> ^a
	Mean	SD	N	Mean	SD	N	
Achievement	15.69	5.55	29	13.52	4.76	29	1.58
Deference	11.62	4.00	29	10.90	3.04	29	.77
Order	11.86	3.95	29	8.83	3.58	29	3.03**
Exhibition	15.93	2.27	29	14.14	4.58	29	— ^b
Autonomy	15.33	4.04	30	11.89	3.84	28	1.46
Affiliation	17.17	3.25	29	14.72	3.87	29	2.58**
Intracception	20.47	4.68	30	15.71	5.32	28	3.61**
Succorance	11.20	4.69	30	10.71	3.68	28	.44
Dominance	16.86	3.60	29	12.97	4.82	29	3.44**
Abasement	12.63	4.72	30	12.46	5.36	28	.13
Nurturance	16.97	4.75	31	12.93	4.61	27	3.26**
Change	17.39	4.65	28	15.43	5.14	30	1.51
Endurance	14.26	4.16	27	11.23	4.31	31	2.71**
Heterosexuality	18.24	4.94	34	15.00	5.01	24	2.36*
Aggression	14.36	4.62	28	9.50	4.47	30	4.05**

^a One-tailed *t* tests were used.^b The Cochran-Cox approximation was used because of heterogeneity of variance.

* Significant at .05 level.

** Significant at .01 level.

statement social desirability. For the 58 subjects the range of endorsement percentage in the individually socially desirable direction was from 35.03 to 79.66, the subject providing the lower end of the range being unique in that he was the only subject to fall below the 50% mark.

Individual vs. group social desirability values. When the group mean social desirability values of the statement alternatives were considered, it was found that the subjects endorsed the group-defined higher socially desirable statement on 55.80% of the 204 imperfectly matched items. This result is very similar to that reported by Goodstein and Heilbrun (1959) who found, for an independent sample of 248 college subjects, that the group-defined more socially desirable statement was endorsed 55.92% of the time on the average. The 55.80% value found in the present study differs significantly ($t = 8.99$ for 57 *df*, $p < .001$) from a chance expectancy of 50%, clearly indicating that performance on the EPPS is also related to group-defined desirability of the statements.

A comparison between the predictableness of EPPS statement endorsement from group social desirability values and from individual

social desirability values was made, and it was found that the difference between the mean percentage of 67.16 for items in which the endorsed statement was the more individually desirable was significantly higher than the 55.80% of items in which the endorsed statement was group-defined more desirable ($t = 9.86$ for 56 *df*, $p < .001$).

Individual social desirability and separate EPPS scales. Although it had been shown in this study that the individual's judged desirability of the statement alternatives does show a significant relationship to statement endorsement on the EPPS, the question still remained as to the extent of this relationship for the separate scales of the test. Data pertinent to this question are presented in Table 1 which gives the means and standard deviations on each of the 15 EPPS scales for the 15 separate high and low social desirability groups and the results of *t* test comparisons. It can be seen that the group of subjects rating the trait statements as more highly desirable had the higher mean on each scale when compared to the group rating the trait statements as less desirable, the differences being significant for 9 of the 15 comparisons and approaching significance at less than the .10

level on 3 more. Thus the positive relationship between individual desirability and endorsement appears to hold for most but not for all of the EPPS scales.

DISCUSSION

One major finding in the present study is that college subjects endorsed the individually defined, more socially desirable response as self-characteristic on at least two out of every three items in which the social desirability values of the paired statements differed. This would seem to clearly indicate that social desirability has not been eliminated as an important source of performance variance on the EPPS as has been suggested by some previous investigators (Borislow, 1958; Kelleher, 1958; Navran & Stauffacher, 1954). The inconsistencies in results among the many investigators bearing upon this issue most likely has stemmed from the use of group social desirability statistics which, based on the results of the present study, are less likely to be related to the individual's response endorsement than the desirability values for that particular individual.

A second major interest in this study was whether an individual's responses to a structured personality questionnaire are more highly related to his own judged social desirability of the responses or to averaged group desirability values. Taylor's (1959) recent finding that group values were more highly related appeared to be inconsistent with the logical assumption that the closer individual parameters of behavior are approximated, the more effectively can behavior be predicted for the individual. The results of the present study were not in agreement with those of Taylor, since it was found that college subjects selected the statement alternative on the EPPS which was individually more desirable more than 67% of the time while these same subjects endorsed the group-defined more desirable statement on less than 56% of the items. Taylor suggested that the lesser relationship obtained between individual social desirability values and MMPI performance than between group values and performance could be attributed in large measure to the lower reliability of individual values which, in turn, would serve to more seriously atten-

uate this correlation. If this is true, it should be pointed out that the opposite results were found in the present study *despite* the lower reliabilities of the individual social desirability values. Since there were so many procedural differences between the Taylor study and the present one (e.g., nature of test item content, true-false vs. forced choice format of the tests, nature of the subject population, etc.,) further speculation regarding the differential results appears unwarranted at the present time.

The analysis of relationships between individual desirability and specific scores suggests a fairly general positive relationship over scales. An interesting finding was that the three scales which appear to be unrelated to social desirability (Deference, Succorance, and Abasement) have a common psychological element: their characteristic behaviors involve some type of subordination of the individual to another person. This suggests the hypothesis that the young adult college subjects used in this study, being in transition from a period of subordinate childhood relationships into a period of more superordinate or coequal adult relationships, have not as yet stabilized their value systems regarding subordinate, equal, and superordinate interpersonal roles. A consequence of this would be that what the subject perceived as socially desirable and self-characteristic relative to role-related behaviors would be at least temporarily unrelated.

SUMMARY

This study investigated three issues relative to performance on the Edwards Personal Preference Schedule: (a) the relationship between the social desirability of statement alternatives on the questionnaire and the endorsement of statements as self-characteristic when the individual's own desirability values are used as predictors, (b) the prediction of statement endorsement using the individual's social desirability values for the statements as opposed to prediction using group-defined values for the statements, (c) the differential relationships between the individual's social desirability values and performances on the separate scales of the EPPS.

The results indicated that the individual's social desirability set is an important source of variance in EPPS performance, since college subjects endorsed as self-characteristic the more highly valued statement at least two out of every three occasions when the statement alternatives were assigned different individual social desirability values. It was also found that individual social desirability values were more highly related to EPPS performance than were group desirability values, a finding which differs from that reported by Taylor (1959) who used the MMPI as the performance variable. Finally, analysis by separate scales suggested that individual social desirability set is related to most but not all of the EPPS variables.

REFERENCES

- BORISLOW, B. The Edwards Personal Preference Schedule and fakability. *J. appl. Psychol.*, 1958, 42, 22-27.
- CORAH, N. L., FELDMAN, M. J., COHEN, I. S., GRUEN, W., MEADOW, A., & RINGWALL, E. A. Social desirability as a variable in the Edwards Personal Preference Schedule. *J. consult. Psychol.*, 1958, 22, 70-72.
- EDWARDS, A. L. *The social desirability variable in personality assessment and research*. New York: Dryden, 1957.
- EDWARDS, A. L. *Personal Preference Schedule*. New York: Psychological Corp., 1959. (a)
- EDWARDS, A. L. Social desirability and personality test construction. In B. M. Bass & I. A. Berg (Eds.), *Objective approaches to personality assessment*. Princeton: Van Nostrand, 1959. (b)
- EDWARDS, A. L., WRIGHT, C. E., & LUNNEBORG, C. E. A note on "Social desirability as a variable in the Edwards Personal Preference Schedule." *J. consult. Psychol.*, 1959, 23, 558.
- GOODSTEIN, L. D., & HEILBRUN, A. B. The relationship between personal and social desirability scale values of the Edwards Personal Preference Schedule. *J. consult. Psychol.*, 1959, 23, 183.
- HEILBRUN, A. B. Relationships between the Adjective Check-List, Personal Preference Schedule, and desirability factors under varying defensiveness conditions. *J. clin. Psychol.*, 1958, 14, 283-287.
- KELLEHER, D. The social desirability factor in Edwards' PPS. *J. consult. Psychol.*, 1958, 22, 100.
- NAVLAN, L., & STAUFFACHER, J. C. Social desirability as a factor in Edwards Personal Preference Schedule performance. *J. consult. Psychol.*, 1954, 18, 442.
- TAYLOR, J. B. Social desirability and MMPI performance: The individual case. *J. consult. Psychol.*, 1959, 23, 514-517.

(Received April 13, 1960)

BEHAVIOR PROBLEMS OF MIDDLE CHILDHOOD¹

DONALD R. PETERSON

University of Illinois

Before the etiology and treatment of children's behavior disorders can be sensibly examined, the disorders themselves must be defined. For the sake of generality and descriptive efficiency, any concepts employed in such definition should be nonarbitrary, unitary, and independent. Factor analytic methods have been employed with salutary effect in the structural definition of adult disorders (e.g., Lorr, Jenkins, & O'Connor, 1955; Rubenstein & Lorr, 1957; Wittenborn, 1951; Wittenborn & Holzberg, 1951), but similar work with the disorders of childhood has only begun (Hewitt & Jenkins, 1946; Himmelweit, 1953). The present study extends and refines this earlier research by factorizing uniformly gathered judgments of problem behavior during the kindergarten and elementary school years, and by examining changes in problem expression during that time.

SUBJECTS AND PROCEDURES

In the absence of any accepted theory of structural organization among children's behavior disorders, a sample of problems was chosen by empirical means. The referral problems of 427 representatively chosen cases at a guidance clinic were recorded, and frequencies tabulated for all problems mentioned more than once. Groups of synonymous terms were reduced by eliminating all but the most frequently used expression, and four concepts were discarded because they were conceptually supraordinate to other terms, and hence redundant. Choice among the

¹ This study was supported by a grant from the Department of Public Welfare, State of Illinois. I am grateful to Jared T. Lyon, Superintendent of the Hoopston (Illinois) Public Schools; to Lester J. Grant, Charlotte Meyer, and Hazel Dunivan of the Decatur Public Schools; to Donald W. Dunn and John Carlino of the Springfield Public Schools; and to Glenn Raymond, Superintendent of Elementary Schools in Watseka, Illinois, for their help in collecting the data. Most of all, I wish to thank the teachers who worked so hard and with such evident care on the ratings themselves.

remaining variables was determined exclusively by the frequency with which they had occurred, and the 58 most common problems were selected for general investigation.

In use, the variables were ordered randomly, assembled in a format requiring ratings of 0 (no problem), 1 (mild problem), or 2 (severe problem), and submitted for completion to 28 teachers of 831 kindergarten and elementary school children in six different schools in Illinois. The choice of school children, rather than clients undergoing treatment for judged disorders, was based on the assumption that most such disorders are extremes of continuous "normal" dimensions, and was determined by the desirability of obtaining uniform data on large numbers of subjects within the age range under consideration. The large sample requirement has been met previously (Hewitt & Jenkins, 1946; Himmelweit, 1953) by recourse to case history information, but the dangers of that expedient seemed greater than those in the present course, and the study was begun in the hope that otherwise unselected school children would present sufficiently numerous, severe problems to warrant sensible analysis and yield meaningful results. Distributions of ratings were generally eccentric, but the effects were reduced by excluding some rarely checked problems (dizziness, soiling, and enuresis, which occurred in less than 3% of the cases, were eliminated), and by pooling judgments of mild and severe problems (ratings of 1 and 2) for all the remaining variables.

For analysis, the sample was divided into four groups: a kindergarten sample ($N = 126$), a first and second grade sample ($N = 237$), a group from the third and fourth grades ($N = 229$), and a fifth and sixth grade sample ($N = 239$). Two teacher ratings were available for each kindergarten child; the number of actual ratings used in the analysis is thus double the N given above for the kindergarten group. Phi coefficients of intercorrelation were computed separately for the four samples. From each correlation matrix, 10 centroid factors were extracted, and from each set of centroid factors 2 were rotated to conform with Kaiser's varimax criterion (Kaiser, 1958).

Human judgment was involved only once in all that analysis—in deciding how many factors to retain for rotation. The decision to keep only two was based on inspection of plots of variance removed by successive centroid factors, and the application of criteria for factor retention developed elsewhere

(Peterson, 1960). Out of personal curiosity, five-factor solutions were also tried for each data set; but these, as expected, were much less stable over age than the two-factor solutions, and only the latter will be reported.

Factor scores were computed for all cases by unweighted summation of pertinent problems checked by the teachers. Interjudge correlations were computed for the kindergarten sample, and further attention directed toward comparing the four age groups. Data for boys and girls were separated in all comparisons, because of well known sex differences in problem expression, and mean factor scores were computed to show trends in the development of behavior problems over the years of middle childhood.

RESULTS

The Factors

All four sets of rotated factor loadings are presented together in Table 1, an arrangement permitted only by the marked similarity between results at the four age levels.² Factor 1 is obviously a *conduct problem* dimension, closely resembling the like-named factor isolated by Himmelweit (1953) and "unsocialized aggression" as defined by Hewitt and Jenkins (1946). Factor 2 has been labeled *personality problem* in accordance with Himmelweit's designation and common usage. It is much like the "over-inhibited behavior" dimension which Hewitt and Jenkins found. Actually these terms, "personality problem" and "conduct problem," are grossly inappropriate. Both problems are personality expressions, and both affect conduct. But the central meanings seem clear enough. In one case, impulses are expressed and society suffers; in the other case impulses are evidently inhibited and the child suffers.

The generality of these factors appears to be enormous. Not only do they emerge with striking uniformity over the limited age range and the particular variables and subjects examined here; they have appeared in very much the same form with the recorded prob-

lems of treatment cases (Hewitt & Jenkins, 1946; Himmelweit, 1953), and remarkably similar factors have appeared in the questionnaire behavior of delinquent boys (Peterson, Quay, & Cameron, 1959). Considering all studies together, age has varied from early childhood to adolescence; problem status has varied from none, through clinic attendance to incarceration for delinquency; data sources have varied from case history records, to standard ratings, to questionnaire responses; methods of factor extraction have varied from cluster inspection to centroid analysis; rotational methods have varied from none, through visual shifts to both orthogonal and oblique solutions, to analytic techniques. Through it all, the factors have stayed the same, and their definition at last seems adequate. The time is ripe for study, particularly experimental study, of dynamics, etiology, and treatment.

Factor Scores and Their Reliability

Such investigations, however, cannot proceed until various properties of the measuring devices have been examined. Factor scores were computed by unweighted summation over the first 15 variables for each factor as listed in Table 1. Below that point, many of the variables either have no appreciable loading on either dimension, or approximately equal loadings on both. The former condition holds especially for skin allergy, hay fever, nausea, and stomach-aches, which may often be purely somatic and qualitatively distinct from the other variables examined. The latter condition, roughly equal loadings on both factors, holds for crying, nervousness, and certain school attitudes, variables which are either very general in nature or exhibit some kind of developmental change.

Reliability and interfactor correlation were examined for the kindergarten group only, since only for that group were dual ratings available. Interjudge r 's of .77 and .75 were found for Factors 1 and 2, respectively. These figures are exceptionally good for ratings, and are sufficiently high for most research purposes. The correlation between factors was .18, low enough to meet most requirements for independence.

² The rating schedule, correlation matrices, and unrotated centroid factor matrices have been deposited with the American Documentation Institute. Order Document No. 6632 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance \$2.00 for microfilm or \$3.75 for photocopies. Make checks payable to: Chief, Photoduplication Service, Library of Congress.

TABLE 1
ROTATED FACTOR LOADINGS

Factor	Conduct Problem				Personality Problem			
	K ^a	1-2	3-4	5-6	K ^a	1-2	3-4	5-6
Conduct Problem								
Disobedience	.74	.77	.69	.86	.03	.04	.07	.11
Disruptiveness	.73	.67	.66	.76	-.04	.19	-.03	.11
Boisterousness	.68	.63	.67	.68	-.16	.07	-.07	-.09
Fighting	.54	.73	.61	.77	-.09	-.04	.11	.07
Attention-seeking	.54	.67	.63	.76	-.12	.10	-.07	.02
Restlessness	.64	.58	.62	.71	.04	.24	.06	.20
Negativism	.56	.64	.60	.70	.12	.27	.20	.15
Impertinence	.57	.57	.53	.76	.02	-.08	.00	.08
Destructiveness	.59	.65	.51	.65	-.05	.27	.19	.00
Irritability	.53	.59	.57	.69	.01	.11	.04	.07
Temper tantrums	.54	.37	.49	.64	.08	.11	.22	.16
Hyperactivity	.51	.49	.54	.49	-.06	.12	.00	.03
Profanity	.30	.42	.64	.60	-.07	.11	.02	.00
Jealousy	.23	.50	.41	.56	.06	.10	.12	.11
Uncooperativeness	.67	.67	.53	.71	.09	.31	.38	.21
Distractibility	.56	.57	.61	.72	.29	.42	.32	.26
Irresponsibility	.60	.65	.49	.65	.22	.18	.47	.20
Inattentiveness	.54	.61	.36	.69	.39	.30	.57	.28
Laziness in school	.44	.59	.36	.37	.29	.36	.55	.31
Shortness of attention span	.48	.54	.31	.60	.37	.34	.55	.29
Dislike for school	.38	.40	.32	.41	.06	.26	.54	.13
Nervousness	.22	.25	.46	.50	.40	.44	.22	.26
Thumb-sucking	.29	.17	.36	.05	.09	.28	-.03	.15
Skin allergy	-.16	.02	-.20	-.05	.01	.21	.05	-.20
Personality Problem								
Feelings of inferiority	.12	.25	.13	.17	.59	.56	.66	.62
Lack of self-confidence	.12	.26	.13	.16	.60	.61	.60	.58
Social withdrawal	-.03	.08	.04	.05	.50	.64	.61	.60
Proneness to become flustered	.07	.28	.15	.24	.54	.59	.60	.58
Self-consciousness	-.13	-.03	-.15	.16	.55	.60	.47	.63
Shyness	-.16	-.18	-.23	-.13	.62	.57	.50	.51
Anxiety	.01	.19	.24	.10	.50	.57	.55	.47
Lethargy	-.06	.22	.01	.31	.52	.47	.61	.43
Inability to have fun	-.15	.06	-.14	-.09	.49	.48	.53	.48
Depression	.00	.20	.04	.29	.47	.43	.64	.42
Reticence	.06	.20	.08	.14	.45	.43	.64	.41
Hypersensitivity	.06	.26	.18	.30	.40	.53	.54	.46
Drowsiness	.02	.09	.09	.29	.39	.48	.45	.41
Aloofness	-.16	-.03	-.04	.05	.51	.32	.50	.31
Preoccupation	.09	.12	.23	.37	.47	.57	.64	.41
Lack of interest in environment	.24	.30	.21	.51	.40	.44	.67	.28
Clumsiness	.16	.21	.36	.17	.43	.54	.34	.36
Daydreaming	.14	.26	.21	.49	.53	.46	.69	.47
Tension	.21	.31	.39	.39	.41	.62	.27	.41
Suggestibility	.04	.29	.31	.52	.41	.42	.48	.30
Crying	.15	.14	.06	.59	.27	.48	.32	.19
Preference for younger playmates	.28	.21	.23	.14	.08	.45	.37	.32
Specific fears	-.09	.24	-.02	-.04	.24	.47	.20	.20
Stuttering	.11	.17	.08	.02	.27	.35	.29	.16
Headaches	.19	.21	.00	.00	.07	.46	.22	.27
Nausea	-.10	.23	.07	-.02	.01	.14	.38	.37
Truancy from school	.27	.07	.04	.22	.00	.20	.39	.35
Stomach-aches	.10	.18	.05	-.06	-.01	.30	.38	.29
Preference for older playmates	-.14	.05	.26	.16	.01	.38	.16	.01
Masturbation	.08	.14	.26	.04	-.18	.40	.04	.17
Hay fever or asthma	.15	-.01	.17	-.11	.05	.21	.09	.03

^a Kindergarten.

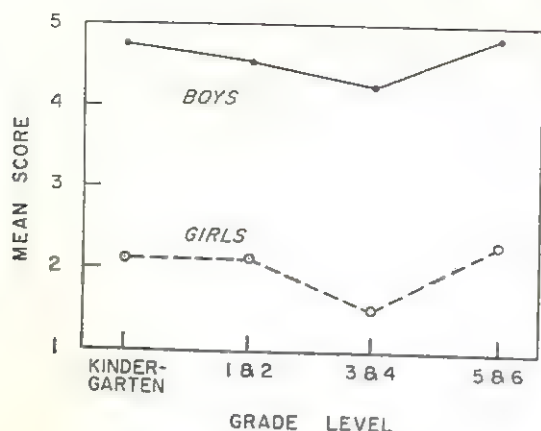


FIG. 1. Mean conduct problem scores.

Developmental Changes

Mean factor scores were computed for boys and girls in all age groups, and the results are shown in Figures 1 and 2. Throughout middle childhood, boys consistently display more severe conduct disturbances than girls, possibly as a function of constitutional differences, but more likely in response to different levels of social expectancy and tolerance for misbehavior. An interesting reversal, however, occurs in the expression of personality problems. Boys evidently start school with more personality problems than girls, but around the seventh or eighth year such problems become more plentiful among girls. Again, social pressures for sex-type conformity seem the likeliest causal agents. Reasons for the apparent upswing in problems at the fifth and sixth grade level are obscure. The increase

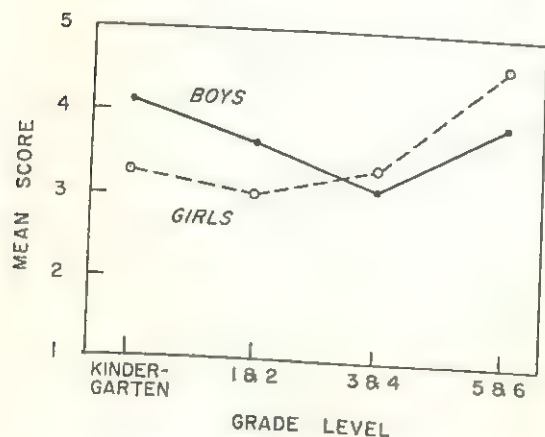


FIG. 2. Mean personality problem scores.

may arise from the early agitation of adolescence, and the difficulty this can bring about in our society.

SUMMARY

This study was designed to improve structural definition of children's behavior problems and to examine changes in those problems over the years of middle childhood. Teacher ratings of 58 clinically frequent problems were obtained for 831 kindergarten and elementary school children, and four separate factor analyses were conducted, one for the kindergarten subjects and one each for children in grades 1-2, 3-4, and 5-6. Two factors emerged with remarkable invariance in all four analyses. The first implied a tendency to express impulses against society, and was labelled "conduct problem." The second contained a variety of elements suggesting low self-esteem, social withdrawal, and dysphoric mood. It was called "personality problem." Both factors have now appeared in a number of studies despite wide differences in subjects, variables, and analytic procedures.

Comparisons over age showed that boys displayed more severe conduct problems than girls at all age levels examined. Kindergarten and primary school boys also showed more severe personality problems than girls, but at the two highest age levels this trend was reversed, and girls displayed more personality problems than boys.

The definition of both dimensions seems adequate. Reliable, independent measures of the factors can be obtained, and the way toward investigation of dynamics, etiology, and treatment now seems clear.

REFERENCES

- HEWITT, L. E., & JENKINS, R. L. *Fundamental patterns of maladjustment: The dynamics of their origin*. Springfield, Ill.: Green, 1946.
- HIMMELWEIT, HILDE T. A factorial study of "children's behavior problems." Cited in H. J. Eysenck, *The structure of human personality*. London: Methuen, 1953.
- KAISER, H. F. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 1958, 23, 187-200.

- LORR, M., JENKINS, R. L., & O'CONNOR, J. P. Factors descriptive of psychopathology and behavior of hospitalized psychotics. *J. abnorm. soc. Psychol.*, 1955, 50, 78-86.
- PETERSON, D. R. The age generality of personality factors derived from ratings. *Educ. psychol. Measmt.*, 1960, 20, 461-474.
- PETERSON, D. R., QUAY, H. C., & CAMERON, G. R. Personality and background factors in juvenile delinquency as inferred from questionnaire responses. *J. consult. Psychol.*, 1959, 23, 395-399.
- RUBENSTEIN, E. A., & LORR, M. Patient types in outpatient psychotherapy. *J. clin. Psychol.*, 1957, 13, 356-361.
- WITTENBORN, J. R. Symptom patterns in a group of mental hospital patients. *J. consult. Psychol.*, 1951, 15, 290-302.
- WITTENBORN, J. R., & HOLZBERG, J. D. The generality of psychiatric syndromes. *J. consult. Psychol.*, 1951, 15, 372-380.

(Received April 18, 1960)

ATTRIBUTION OF TRAITS AND EMOTIONAL HEALTH AS FACTORS ASSOCIATED WITH THE PREDICTION OF PERSONALITY CHARACTERISTICS OF OTHERS¹

MARVIN SPANNER²

University of California, Berkeley

Many unsystematized explanations have been offered concerning the manner in which judges predict the personality characteristics of others. The present paper will focus on two theories currently popular. The first, based on an interpersonal theory of personality, has been most extensively represented by Harry Stack Sullivan. Sullivan's (1947) ideas on the subject are reflected in his famous dictum:

"It is not that as ye shall judge so shall ye be judged, but as you judge yourself, so shall you judge others; strange but true so far as I know, and with no exception."

The above quotation suggests that interpersonal prediction is based primarily on an attributive mechanism in which the state of the individual's self-concept determines the quality of the interpersonal appraisal. One

¹ This paper is based in part on a thesis submitted in partial fulfillment of the requirements for the PhD in the Department of Psychology of the University of California, Berkeley. The writer wishes to express his appreciation and indebtedness to Harrison G. Gough, thesis advisor; Donald W. MacKinnon, director of the Institute of Personality Assessment and Research; and Victor B. Cline.

This research was supported in part by the United States Air Force under Contract No. AF 18 (600)-8, monitored by Technical Director, Detachment No. 7 (Officer Education Research Laboratory), Air Force Personnel and Training Research Center, Maxwell Air Force Base, Alabama. Permission is granted for reproduction, translation, publication, use, and disposal, in whole and in part, by or for the United States Government. Personal views or opinions expressed or implied in this publication are not to be construed as necessarily carrying the official sanction of the Department of the Air Force or of the Air Research and Development Command.

² Now at Neuropsychiatric Institute, University of California, Los Angeles.

can deduce from his statement the following hypothesis:

I. There is a positive relationship between accuracy in predicting the personality characteristics of others and similarity of personality characteristics of the judge and the individual being judged. This would logically appear to follow because an understanding of others is based, according to Sullivan, on one's self. Thus, if one tends to perceive others as similar to one's description of one's self, then those actually similar will be perceived accurately. Lingren and Robinson (1953) make a similar point when they state, in criticizing Dymond's study (1950), that "Conventional people get good scores on empathy tests because most of their partners (or referents) in the test are also conventional."

The second group of explanations of predictive ability involves the mental health of the individual. Cline (1953) in an exhaustive review of the literature suggests that, on the whole, a good judge of others is emotionally sound, has good interpersonal relations, is happier, more popular and flexible. In addition, most clinicians assume that an emotionally healthy individual is a better judge of personality because he has fewer problems which interfere with his understanding of others. In the present instance, two measures have been utilized to evaluate the emotional health of the subjects. The first is the correlation between the judge's evaluation of himself and his evaluation of the sort of individual he would ideally like to be. The second is a composite rating by experienced observers of the "soundness" of the judge.

We have, therefore, the following two hypotheses relating the "mental health" of the judge and his ability to predict personality characteristics of others:

II. There is a positive relationship between the self-ideal-self correlation of the judge (called self-acceptance and used as a measure of mental health) and his ability to predict personality characteristics of others.

III. There is a positive relationship between the rated "soundness" of a judge (used as a measure of mental health) and his ability to predict personality characteristics of others.

METHOD

Since both the tests of judging ability and the prediction instruments were devised and fully reported by Cline (1955), the following is an abbreviated description of these procedures.

Tests of Judging Ability

The basic procedure used was a sound movie of four fictitious employment interviews. The interviewees were male undergraduates. Three of them were 19 years old and single. The fourth was 33 years old, married, childless, and a veteran. Each observer or judge was asked, after viewing each one of the four interviews, to make a series of predictions about the interviewee (hereafter called social object) on three specially devised instruments, only one of which was used in the present analysis.

The four filmed interviews, conducted by a trained actor, were highly structured and fairly constant. These were chosen from a group of nine films in order to obtain the most diverse personalities among the social objects. Cline has stated that the interviews are divided into three phases: a standard job interview situation, a stress situation in which the interviewer is highly critical of the interviewee, and a relaxed after-interview abreaction session in which the interviewee's reactions to the interview are discussed. The intent was to get a dynamic record of a series of subjects on a sound film, rich and varied enough in the use of verbal and visual cues so that a judge would have an adequate sample of each social object's social technique.

Prediction Task

The sole prediction task used in the present study to assess the judging ability of persons who observed the films was the Personality Word Card (PWC). Here the judge was asked to predict the responses of each of the four social objects on a 100-word adjective check-list. This list was derived from a 300-word Adjective Check-List compiled by Gough.³

³ Gough, H. G. Predicting success in graduate training. (Progress report) Unpublished manuscript, University of California Institute of Personality Assessment and Research, 1952.

The corrected split-half reliability for the PWC as a test of judging ability for a group of 100 college undergraduates was $.83 \pm .06$.

The PWC was also used by both the judges and social objects to describe themselves. Comparisons were, therefore, readily made among the judges' prediction of the social object, the social objects' self-reports, and the judges' self-reports.

Two other prediction tasks were given to the judges, although they were not used in the present analysis. One of them, the Behavioral Postdiction Test, in which the judge was required to predict the social object's behavior in real life, was discarded because it was discovered, when a simple analysis of variance was carried out, that the source of variation attributable to the judges was not significant. A third instrument, a multiple-choice sentence completion test, was also discarded as a test of judging ability, because Cline had found that the corrected split-half reliability for a group of 100 college undergraduates was only $.36 \pm .06$.

Judges

One hundred Air Force Captains were tested at the Institute of Personality Assessment and Research, as part of a large scale study of officer effectiveness. They were tested in groups of 10 in a living-in situation similar to that employed by the OSS during World War II (Office of Strategic Services, 1948). Their average age was 33.6 with a range of 27 to 49. Ninety-five percent of the officers were married, with an average of two children. Fifty-six had attended college, seven had done some graduate work, and only three had not graduated from high school.

Measures

Accuracy. This variable is defined as the tetrachoric correlation between the social objects' self-reports and the judges' prediction of them on the PWC. Since there are four social objects, each one of the 100 judges has four accuracy scores.

Similarity. This measure is defined as the tetrachoric correlation between the social objects' self-reports and the judges' self-reports on the PWC. Here too, each one of the 100 judges has a similarity score for each of the four social objects.

Self-Acceptance (self-ideal-self correlation). Each judge completed the PWC twice, once describing himself as he was and then a second time describing himself as he would ideally like to be. The measure of association used was the phi coefficient. It is assumed that the greater the degree of association the more self-accepting and therefore the "healthier" the individual.

Soundness. This variable was drawn from a pool of 30 traits which were used by 10 staff members in rating the judges (officers). The following procedure was used. After the first four groups of 10 officers had been run, 10 staff members of the Institute of Personality Assessment and Research rated each of the 40 officers in the 30 traits. A normal distribution was required of the raters, using

TABLE 1
CORRELATION COEFFICIENTS (PEARSONIAN) BETWEEN
ACCURACY AND SIMILARITY, SELF-ACCEPTANCE, AND
SOUNDNESS, FOR EACH OF THE FOUR
SOCIAL OBJECTS

Measure	Accuracy in Predicting the Four Social Objects			
	I	II	III	IV
Similarity	.01	.28*	.00	.00
Self-Acceptance	-.06	.15	.06	-.03
Soundness	-.10	-.13	.13	-.09

* $p < .01$.

a five-point scale. The same procedure was followed in rating the remaining 60 judges, after they had all been assessed. A composite trait rating was derived from each assessee by combining the individual ratings of the 10 staff members. Each item of the trait pool was then assigned a scale value for each assessee.

Soundness was defined as:

Maturity in personal relations; self-insight and self-acceptance, as well as acceptance and understanding of others. Absence of serious emotional problems. Stability of mood and manner. Good balance of social conformity and spontaneity.

All of the above measures, once obtained, are simply treated as scores, for purposes of statistical analysis.

RESULTS

The first hypothesis—that accuracy of prediction of personality traits of others is dependent on the degree of similarity of judge and social object—was tested by correlating similarity and accuracy for each of the social objects. As can be seen in Table 1, the correlation between similarity and accuracy for Social Object I is .01 and for Social Objects III and IV is .00. The correlation for Social Object II, however, is .28, significant at the .01 level of confidence. Since this is the only one of the four social objects where significance was achieved, the first hypothesis, that there is a relationship between accuracy of prediction and similarity of judge and social object, was not substantiated.

The following results are concerned with the relationship of mental health and the ability to predict personality characteristics of others. As can be seen in Table 1 there is no significant relationship between the ability

to predict personality characteristics of others and self-acceptance (self-ideal-self correlation) or soundness of the judge, and, therefore, both Hypotheses II and III have not been substantiated.

Although it has been shown that similarity of personality characteristics of judge and social object is not related to accuracy of prediction on the part of the judge, other factors might prevent a simple expression of this relationship. For example, a judge who is similar to a social object may predict accurately only if he is self-accepting (i.e., his self-ideal-self correlation is high). With this in mind, a triple classification analysis of variance for judging ability was carried out. The judges were divided into a high and low similarity group as well as a high and low self-acceptance group (self-ideal-self correlation), for each of the four social objects. Judges above the median were considered high and those below the median were considered low on both similarity and self-acceptance.

Because of the varying number of observations in the cells of the triple classification, it was necessary to pool the residual and triple interaction sums of squares. This is a conservative estimate of the error variance; if the triple interaction is significant, pooling it with the error variance would only produce an overestimate of this term and tend to produce smaller F ratios.

TABLE 2
ANALYSIS OF VARIANCE OF JUDGING ABILITY, IN TERMS
OF SIMILARITY AND SELF-ACCEPTANCE, FOR THE
FOUR SOCIAL OBJECTS

Source of Variation	SS	df	MS	F
Similarity (Sim)	656	1	656	3.14
Self-Acceptance (SA)	167	1	167	.80
Social Objects (SO)	92,219	3	30,740	147.08*
Sim \times SO	957	3	319	1.53
SA \times SO	761	3	254	1.22
Sim \times SA	3,693	1	3,693	17.67*
Sim \times SA \times SO and Residual	80,836	387	209	
Total	179,289	399		

* $p < .01$.

TABLE 3
ACCURACY AND "FAVORABLE" ADJECTIVES
FOR EACH OF THE SOCIAL OBJECTS

Social Object	Mean Accuracy	Percentage of Gough's "Favorable" Adjectives Used by Social Objects in Their Self-description
I	46.24	42
II	23.96	30
III	33.51	31
IV	64.51	48

The analysis of variance shown in Table 2 reveals two significant mean squares: social objects, and the interaction between similarity and self-acceptance. In attempting to understand the basis for the large mean square for social objects, it was noted that the self-descriptions of the social objects tended to be quite favorably toned. A measure of this favorable tone was derived from Gough's list of "favorable" adjectives. Gough (see Footnote 3) was able to derive the list by asking a group of 30 college students to rate the entire list of adjectives for its favorability. The highest rated 25% were selected for the favorability key. One question which immediately comes to mind is whether there is a relationship between accuracy of prediction, on the part of the judge, and favorableness of self-perception, on the part of the social object. If one compares the list of favorable adjectives with the mean accuracy score for each of the social objects, in Table 3, one notes a perfect correlation (ρ) between the two variables. Thus, the accuracy of prediction appears to depend, in part, on favorableness of self-perception of the social object.

The second significant mean square noted in Table 2 is the interaction between similarity and self-acceptance. To further clarify the relationship, the accuracy scores of judges conforming to the four possible combinations of similarity and self-acceptance, under the present conditions, are shown in Table 4.

Two groups can be considered relatively accurate: judges similar to a social object and self-accepting, and judges dissimilar to a social object and not self-accepting. The other two groups can be considered relatively inaccurate: judges similar to a social object and not self-accepting, and judges dissimilar to a social object and self-accepting.

DISCUSSION AND CONCLUSIONS

What are the implications of Sullivan's statement that one judges others only in terms of one's self? It obviously creates a huge solipsism in which no distinction is drawn between one's self and others in the judgment process. The lack of relationship exhibited between similarity of judge and social object and accuracy of prediction (Hypothesis I), in the present experiment, suggests that the prediction of personality characteristics of others is not a reflection of the judge, a simple attribution to another of the traits of the judge.

In addition, no relationship was found, in the present study, between the mental health of an individual—using (a) a self-rating criterion, self-acceptance (self-ideal-self correlation); and (b) an external criterion, judged soundness of an individual by expert raters—and his ability to predict personality characteristics of others (Hypotheses II and III).

When, however, both similarity and self-acceptance (self-ideal-self correlation) interact, there is a significant effect on accuracy. This result can be understood if we make some assumptions concerning the psychological meaning of each of these variables. Let

TABLE 4
MEAN ACCURACY OF THE INTERACTION ON HIGH AND LOW SCORES ON BOTH SIMILARITY AND SELF-ACCEPTANCE

Group	N	Mean	SD
High Sim and High SA	120	46.18	20.07
High Sim and Low SA	80	39.08	22.40
Low Sim and High SA	80	37.49	21.23
Low Sim and Low SA	120	42.97	21.93

Note.—High = Judges above median on Sim or SA; Low = judges below median.

us assume that a judge who is relatively self-accepting will tend to judge others as similar to himself. The unconscious basis for his predictions might be as follows: "I am a nice person and so are other people." A judge who is relatively low on self-acceptance, however, will tend to judge others as dissimilar to himself. Here, the unconscious basis for his predictions might be: "I am not a nice person, but others are." The mechanisms described here are similar to those characterized by Cameron and Magaret (1951) as, respectively, assimilative and disowning projection.

As was described above, there are objective measures of the similarities of judge and social object. If we focus our attention on those judges who are high on self-acceptance, we can hypothesize that they would be accurate in the prediction of personality characteristics of those social objects similar to themselves, and inaccurate in the prediction of personality characteristics of those social objects dissimilar to themselves. This would logically follow since, as postulated above, a self-accepting judge would tend to perceive others as similar to himself. Those actually similar to him would, therefore, be judged accurately, while those dissimilar to himself would be judged inaccurately.

In a similar fashion, we would predict that judges with low self-acceptance scores would accurately perceive those social objects who are dissimilar to themselves, while inaccurately perceiving those social objects similar to themselves. This, too, would logically follow since, as postulated above, judges with low self-acceptance scores would tend to perceive others as dissimilar to themselves. Those social objects actually similar to these judges would, therefore, be perceived inaccurately, while those dissimilar to them would be perceived accurately.

The post hoc explanation invoked here conforms to the results in Table 4. While the results presented suggest that judging the personalities of others involves more than an attribution of one's traits to another, it also suggests that judges engage in a mechanical type of prediction based on their self-concept. If the judges are self-accepting, they judge others in a similar manner. If the judges are

not self-accepting, they judge others in a contrary manner. Accuracy, therefore, depends not only on whether an individual is self-accepting or not, but also on whether the judges are similar or dissimilar to the social object. Bronfenbrenner (1958) has made a similar point. He utilized the variable "Favorability toward others" in his analysis, however, rather than self-acceptance used in the present study.

Looked at from this point of view, the ability to predict personality characteristics of others should not be perceived as a trait residing within the judge, but, rather, as based on the personality patterns of both judge and social object. This point of view is further substantiated by an analysis of variance of judging ability carried out by the author (Spanner, 1955), in which the interaction between judge and social object approached significance at the .05 level.

This approach suggests, therefore, that the question to be asked will no longer exclusively be "Is he a good judge of personality?" but also "Which social objects can he judge well?" Another facet of the same problem would be an attempt to break down the evaluation of personality into various areas and to assess the ability of judges in each area. This presupposes a representative design of the sort suggested by Brunswik (1947), in which there would be not only representative sampling of subjects (judges), but also of objects (social objects). In the present instance a preliminary attempt was made to analyze the different skills involved in judging social objects with favorable as compared to unfavorable self-perceptions. The attempt was abandoned, however, because of the small number of social objects (four).

Possibly if some of the aforementioned approaches are handled successfully, eventually, we will then be in a position to answer the following question: "How does the ability of a judge to predict the personality characteristics of others relate to his own personality dynamics?" This, of course, is related to the entire area of ego psychology and the differential utilization of ego defenses by an individual in judging the personality characteristics of others.

SUMMARY

One hundred military officers were presented with sound movies of stress interviews with four interviewees (social objects) and asked to predict the responses of the social objects on an adjective check-list, as well as to describe themselves on the same instrument. From these basic data measures of accuracy of prediction and similarity of judge and social object were obtained. Measures of the self-acceptance (self-ideal-self correlation) and "soundness" of the judges were also obtained.

Contrary to expectations, accuracy was found to be unrelated to similarity of judge and social object, to self-acceptance, and to the rated soundness of the judge. Accuracy scores were related, however, to a combination of similarity and self-acceptance variables. The concepts of disowning and assimilative projection were utilized as a post hoc explanation of accuracy.

REFERENCES

BRONFENBRENNER, U. The measurement of skill in social perception. In D. C. McClelland, A. L.

- Baldwin, U. Bronfenbrenner, & F. L. Shadtsbeck (Eds.), *Talent and society*. Princeton: Van Nostrand, 1958.
- BBUNSWIK, E. Systematic and representative design of psychological experiments. *U. Calif. Syllabus Ser.*, 1947, No. 304.
- CAMERON, N., & MAGARET, ANN. *Behavior pathology*. Boston: Houghton Mifflin, 1951.
- CLINE, V. B. The assessment of good and poor judges of personality using a stress interview and sound film technique. Unpublished doctoral dissertation, University of California, Berkeley, 1953.
- CLINE, V. B. Ability to judge personality assessed with a stress interview and sound-film technique. *J. abnorm. soc. Psychol.*, 1955, **50**, 183-187.
- DYMOND, ROSALIND F. Personality and empathy. *J. consult. Psychol.*, 1950, **14**, 343-350.
- LINGREN, H. C., & ROBINSON, JACQUELINE. An evaluation of Dymond's test of insight and empathy. *J. consult. Psychol.*, 1953, **17**, 172-176.
- OFFICE OF STRATEGIC SERVICES, Assessment Staff. *Assessment of men*. New York: Rinehart, 1948.
- SPANNER, M. Similarity, identification, and distortion as factors in the prediction of personality characteristics. Unpublished doctoral dissertation, University of California, 1955.
- SULLIVAN, H. S. *Conceptions of modern psychiatry*. Washington, D. C.: William Alanson White Psychiatric Foundation, 1947.

(Received April 18, 1960)

HOSTILITY EXPRESSION AMONG DELINQUENTS OF MINORITY AND MAJORITY GROUPS

DON L. SWICKARD AND BERNARD SPILKA

University of Denver

The frustration-aggression hypothesis has frequently provided a theoretical framework for explaining the phenomena of crime, delinquency (Dollard, Doob, Miller, Mowrer, Sears, Ford, Hovland, & Sollenberger, 1939), and prejudice (Allport, 1954; Dollard et al., 1939). Two prevalent and pervasive sources of frustration which are seen as motivating anti-social behavior are low socioeconomic status and minority group membership (Dollard et al., 1939). Evidence to support this position may be adduced from the work of Glueck and Glueck (1950) and Sutherland and Cressy (1955), who show clearly the relationship between poverty and delinquency. On the other hand, Hammer (1953), McCary (1950), and Mussen (1953) have demonstrated that minority group members reveal more manifest aggression than do members of majority groups. The factors of poverty and discrimination have been confounded in these studies, thus failing to clarify the role of each. It may be hypothesized, however, that minority group membership and low socioeconomic status combine to effect a greater manifestation of hostility than would simply result from the frustrations of poverty alone. Delinquents who are members of a minority might, therefore, be expected to reveal more evidences of hostility than delinquents who are majority group members. The present investigation was designed to assess this position, utilizing Spanish-American and non-Spanish, white delinquents. That the Spanish-American group is a negatively valued minority has been pointed out by Jones (1948) and Saunders (1954) in their writings concerning the Spanish- or Mexican-American, and his

status among Anglo-Americans. Specifically, the hypotheses tested were:

1. A group of Spanish-American delinquents of low socioeconomic status will show significantly greater evidences of aggression, in both amount and kind, than will be demonstrated by a similar group of non-Spanish, white delinquents.
2. Because of tendencies on the part of delinquents to give a "good impression" on personality measures (Lindzey & Goldwyn, 1954; Vane, 1954), it will be necessary to correct hostility measures for social desirability.
3. Social desirability will relate negatively to manifest and extrapunitive measures of hostility and positively to intropunitive and impunitive measures of hostility.

METHOD

Subjects. The original group studied consisted of 81 male and female delinquents on probation at the Denver Juvenile Court. The subjects were selected in such a way as to eliminate all who were the result of mixed group backgrounds, who had been diagnosed as neurotic or psychotic, were other than lower class, or who had been institutionalized in the last 2 years. On the basis of obtaining scores of 10 or more on the MMPI Lie scale, 7 subjects were eliminated. The final group thus consisted of 25 Spanish-American males, 12 Spanish-American females, 25 non-Spanish white males, and 12 non-Spanish white females. All subjects were between 14 and 17 years of age.

Tests. The tests administered were the Rosenzweig & Picture Frustration Study (Rosenzweig, Fleming, & Clarke, 1947), the Siegel Manifest Hostility scale (Siegel, 1956), the 39-item Social Desirability scale, extracted from the MMPI by Edwards (1957), and the MMPI Lie scale (Hathaway & McKinley, 1951).

Procedure. All subjects were tested in groups of from 5 to 15, in the courtroom of the Denver Juvenile Court. All subjects were assured that the testing had nothing to do with their probation and

that the results would not be made available to their probation counselors. All subjects were cooperative and answered all questions without difficulty.

RESULTS AND DISCUSSION

Table 1 contains the means and standard deviations for the four groups on the various tests administered. Pearson product-moment correlations were computed between the Social Desirability scale and the other measures, as seen in Table 2. Hypotheses 2 and 3 obtained support as the Social Desirability scale correlated $-.655$ with the Siegel Manifest Hostility scale, $-.372$ with the extrapunitive scores, $.222$ with the intrapunitive scores, and $.262$ with the impunitive scores. The first two coefficients are significant at the .01 level of confidence, and the last two at the .05 level, with 72 degrees of freedom. One-way simple classification analyses of

variance and covariance were computed between the four groups: Spanish-American males and females, and non-Spanish white males and females. This analysis was more conservative than a group \times sexes design would have yielded. The latter was computed in the analysis of variance for a 2×2 disproportionate subclass numbers design as recommended by Snedecor (1956), in order to check for further significance than might have been obtained via the single variable design. No additional significance was obtained for the group, sex, or interaction terms. Because of the problem of interpreting an analysis of covariance for the 2×2 design with unequal and disproportionate subclass numbers, it was decided to employ the more conservative analysis reported here.

The analyses of variance across groups were significant at the .05 level for social

TABLE 1
MEANS AND STANDARD DEVIATIONS OF AGE, SOCIAL DESIRABILITY, AND MANIFEST HOSTILITY
AND OF EXTRAPUNITIVE, INTRAPUNITIVE, AND IMPUNITIVE SCORES FOR ALL GROUPS

Variable ^a	Group ^b				Total (<i>N</i> = 74)
	SA-M (<i>N</i> = 25)	SA-F (<i>N</i> = 12)	NSW-M (<i>N</i> = 25)	NSW-F (<i>N</i> = 12)	
Age					
Mean	15.0	14.7	15.0	14.5	14.8
SD	.72	.75	.80	.76	.76
SD					
Mean	26.2	20.2	23.6	24.4	24.0
SD	5.10	4.88	6.18	3.59	5.62
MH					
Mean	22.2	23.4	19.6	19.3	21.1
SD	6.62	6.49	8.01	9.14	7.71
E					
Mean	8.5	9.2	8.2	9.0	8.6
SD	2.99	3.96	3.58	3.20	3.42
I					
Mean	7.1	6.4	6.5	6.6	6.7
SD	2.33	2.17	2.44	2.59	2.40
M					
Mean	7.1	6.5	7.7	7.0	7.2
SD	2.23	3.34	2.77	1.58	2.58

^a SD refers to Social Desirability scale; MH, Manifest Hostility scale; E, Extrapunitive Scores; I, Intrapunitive Scores; M, Impunitive Scores.
^b SA-M refers to Spanish-American males; SA-F, Spanish-American females; NSW-M, non-Spanish, white males; NSW-F, non-Spanish, white females.

TABLE 2

CORRELATIONS BETWEEN THE SOCIAL DESIRABILITY SCALE AND THE MANIFEST HOSTILITY SCALE, AND THE EXTRAPUNITIVE, INTRAPUNITIVE, AND IMPUNITIVE SCORES ON THE ROSENZWEIG PICTURE-FRUSTRATION STUDY

Variable	Group				
	SA-M (N = 25)	SA-F (N = 12)	NSW-M (N = 25)	NSW-F (N = 12)	Total ^a (N = 74)
SD and MH	-.648**	-.627*	-.706**	-.666*	-.655**
SD and E	-.110	-.581	-.466	-.505	-.372**
SD and I	-.136	.421	.288	.581*	.222*
SD and M	.104	.588*	.772**	.308	.262*

^a The *r*'s for the total group are not based on an averaging of the subgroup coefficients but on direct computation from the basic data. All subgroup coefficients in the first three rows are homogenous. Those in Row 4 are heterogenous and the total *r* of .262 is at best a conservative estimate of this relationship.

* Significant at .05 level.

** Significant at .01 level.

desirability (see Table 3), and not significant for the Manifest Hostility scale (see Table 3). No significance was obtained for the analyses of extrapunitiveness, intropunitiveness, and impunitiveness (see Table 3). Because of the significant differences among the groups in social desirability and the correlations noted above, analyses of covariance were computed among the groups on the hostility scores, while adjusting the means for social desirability. Significance was attained for the Manifest Hostility scores, but not for any of the other measures (see Table 3). The adjusted means computed for the Manifest Hostility scale are: Spanish-American males, 24.3; Spanish-American females, 19.8; non-Spanish white males, 19.3; and non-Spanish white females, 19.7.

It is apparent that Hypothesis I is partially supported since the Spanish-American male

group obviously manifested more hostility on the Manifest Hostility scale than did any of the other groups.

It is not clear why the Spanish-American females did not reveal more evidences of hostility than did either of the non-Spanish white groups. It is noteworthy that the significance obtained among the groups on the Social Desirability scale is probably, in part, due to the fact that the Spanish-American females received the lowest mean score on this variable.

Jones (1948) in a study of the ethnic patterns of the Mexican family in the United States indicates that the home training given girls in this group is radically different from that given boys. Among other things manifestations of hostility are strongly disapproved. The boys of this group might thus be more similar to the non-Spanish white group in background, and it is possible that the Social Desirability scale is thus less appropriate for the Spanish-American female than for the male. This somewhat different conception of social desirability plus pressures against evidencing overt aggression would account partially for the low Social Desirability scale scores obtained by the girls, and failure of the adjusted Manifest Hostility scores to increase along with those of the Spanish-American males.

The failure of the Rosenzweig Picture-Frustration Study to yield significance has been noted before (Lindzey & Goldwyn, 1954;

TABLE 3

F RATIOS BETWEEN GROUPS FOR ORIGINAL MEANS AND MEANS ADJUSTED FOR SOCIAL DESIRABILITY

Variable	Original Means (df = 3.70)	Adjusted Means (df = 3.69)
SD	3.327*	~
MH	1.011	3.240*
E	<1	<1
I	<1	<1
M	<1	<1

* Significant at .05 level.

Vane, 1954), and it may be that the P-F study is actually not measuring manifest hostility but rather aggression on a different level of the personality than that which might be significant for the present study. The P-F study does, however, seem sensitive to the variable of social desirability, as is the Siegel Manifest Hostility scale, and it is suggested that future research involving these two instruments take this into consideration, as was done here.

Another meaningful consideration may be based on the possibility that the lower-class culture simply supports the expression of hostility without it necessarily being related to frustration. The assumed frustration-aggression relationship may therefore be inappropriate. This would account for the differences observed on the Manifest Hostility scale and the failure of the P-F study to reveal the hypothesized tendencies since the latter measure assumes the significance of frustration.

In brief, the results of the present investigation indirectly lend support to the frustration-aggression hypotheses by suggesting that the combination of frustration from low socioeconomic status and minority group membership may increase the expression of aggression over that which is observed when only low socioeconomic conditions are present.

SUMMARY AND CONCLUSIONS

The present study was designed to investigate the relation of hostility to a combination of low socioeconomic status and minority group membership. The frustration-aggression hypothesis was employed as a theoretical referent, and it was hypothesized that the increased frustration of minority group membership in addition to low socioeconomic status, would produce more manifestations of hostility than would be observed in majority group members of similar class levels. It was further hypothesized that tendencies to give a "good impression" would relate negatively to manifest hostility and extrapunitive tendencies while relating positively to intro-punitiveness and impulsive expressions. On the basis of the latter considerations, it was felt that obtained hostility scores would have to be adjusted to minimize such biases.

Eighty-one Spanish-American and non-Spanish white delinquents on probation served as subjects. All subjects were administered the Siegel Manifest Hostility scale, the Social Desirability scale derived by Edwards, the Rosenzweig Picture-Frustration Study, and the Lie scale from the MMPI. Seven subjects who obtained scores of 10 or more on the Lie scale were eliminated from the study. Significant negative correlations were found between the Social Desirability scale and the Siegel Manifest Hostility scale, and the extrapunitive scores from the Rosenzweig Picture-Frustration Study. Significant positive correlations were obtained between the Social Desirability scale and the measures of intro-punitiveness and impunitiveness. Once the hostility means were adjusted to remove the effects of social desirability, significance was obtained between the groups on the Manifest Hostility scale. The Spanish-American male group was shown to manifest significantly greater hostility on this measure than any other group, thus partially supporting the main hypothesis.

REFERENCES

- ALLPORT, G. W. *The nature of prejudice*. Cambridge: Addison-Wesley, 1954.
- DOLLARD, J., DOOB, L. W., MILLER, N. E., MOWRER, O. H., SEARS, R. R., FORD, G. S., HOVLAND, C. I., & SOLLENBERGER, R. T. *Frustration and aggression*. New Haven: Yale Univer. Press, 1939.
- EDWARDS, A. L. *The social desirability variable in personality assessment*. New York: Dryden, 1957.
- GLUECK, S., & GLUECK, ELEANOR. *Unraveling juvenile delinquency*. New York: Commonwealth Fund, 1950.
- HAMMER, E. F. Frustration-aggression hypothesis extended to socio-racial areas: Comparison of Negro and white children's H-T-P's. *Psychiat. Quart.*, 1953, 27, 597-607.
- HATHAWAY, S. R., & MCKINLEY, J. C. *Manual for the Minnesota Multiphasic Personality Inventory*. (Rev. ed.) New York: Psychological Corp., 1951.
- JONES, R. C. Ethnic family patterns: The Mexican family in the United States. *Amer. J. Sociol.*, 1948, 53, 450-453.
- LINDZEY, G., & GOLDWYN, R. M. Validity of the Rosenzweig Picture-Frustration Study. *J. Pers.*, 1954, 22, 519-547.
- MCCARY, J. L. Ethnic and cultural reactions to frustration. *J. Pers.*, 1950, 18, 321-326.
- MUSSEN, P. H. Differences between the TAT responses of Negro and white boys. *J. consult. Psychol.*, 1953, 17, 373-376.

- ROSENZWEIG, S., FLEMING, E. E., & CLARKE, HELEN JANE. Revised scoring manual for the Rosenzweig Picture-Frustration Study. *J. Psychol.*, 1947, 24, 165-208.
- SAUNDERS, L. *Cultural differences and medical care*. New York: Russell Sage Foundation, 1954.
- SIEGEL, S. M. The relationship of hostility to authoritarianism. *J. abnorm. soc. Psychol.*, 1956, 52, 368-372.
- SNEDECOR, G. W. *Statistical methods*. Ames, Iowa: Iowa State College Press, 1956.
- SUTHERLAND, E. H., & CRESSY, D. R. *Principles of criminology*. Philadelphia: Lippincott, 1955.
- VANE, JANET R. Implication of the performance of delinquent girls on the Rosenzweig Picture-Frustration Study. *J. consult. Psychol.*, 1954, 18, 414.

(Received April 18, 1960)

THE USE OF OPPOSITE SEX SCALES AS A MEASURE OF PSYCHOSEXUAL DEVIANCY¹

B. G. ROSENBERG, B. SUTTON-SMITH

Bowling Green State University

AND E. MORGAN

Monroe County Schools

Most self-report tests of personality obtain an estimate of male and female sex role identification through the use of like-sex scales. Sex role identification or the lack of it is typically evaluated by the subject's responses to a scale of items constructed and validated on his own sex (Hathaway & McKinley, 1943; Strong, 1943). Thus, a person obtaining an average or high score on his own sex scale is said to be identified with his own sex, and a person obtaining a low score on his own sex scale is said to lack identification with his own sex, or to be more like the opposite sex in his responses. In an earlier paper the authors described the derivation of a scale for boys and girls between the ages of 8 and 12 years in which each subject received a score on an opposite-sex scale as well as on a like-sex scale (Rosenberg & Sutton-Smith, 1959). The existence of these scales for measuring masculinity and femininity has made it possible to investigate the relative effectiveness of both the like-sex scale and the opposite-sex scale for discriminating sex role identification.

It is generally assumed that individuals who are faulty or aberrant in their sex role identification will be more emotionally disturbed than those who are not (Henry, 1948; Terman & Miles, 1936). Therefore, they should tend to have less satisfactory scores on measures of emotional stability. This paper examines the relationship between various types of scores on a masculinity and a femininity scale and several independent measures of emotional stability. The independent

measures of emotional stability make it possible to explore not only the existence of sex role confusion, but also certain of its qualitative aspects.

METHOD

A group of 337 children in the fourth, fifth, and sixth grades in two elementary schools in Northwest Ohio and Southeast Michigan comprised the sample.² Several instruments were administered to the entire group: (a) a check list of games and play activities which is described elsewhere (Rosenberg & Sutton-Smith, 1959) and which effectively yields measures of masculinity and femininity, (b) an empirically derived scale which has been found to measure extremes of impulsive behavior (Sutton-Smith & Rosenberg, 1959), (c) the Children's Manifest Anxiety scale which affords a measure of anxiety (Castaneda, McCandless, & Palermo, 1956), and finally (d) the Brown Personality Inventory (Brown, 1934). The latter scale, derived in 1934, has been found to discriminate neuroticism in children, and to be highly correlated with the CMA scale (Rosenberg, Sutton-Smith, & Morgan, in press). Its addition provides subscores which reveal the source of major concern for neurotic children (home, school, physical, insecurity, and irritability). The "home" subscale includes feelings of parental rejection, parental severity, and sibling jealousy. The "school" subscale identifies feelings of inadequacy in the classroom. The "physical" subscale deals with disturbances in physical well-being which appear to be psychogenic in origin. The "insecurity" subscale contains items reflecting anxiety, interpersonal inadequacy, and generalized neurotic concerns. The "irritability" subscale examines emotional reactivity, low tension-binding qualities, and severe feelings of rejection (Brown, 1934).

An analysis was conducted of the scores on impulsiveness, anxiety, and neuroticism of boys and girls in the upper and lower quartiles on the masculinity and femininity scales.

¹ This study was facilitated by a grant from the Scholarly Advancement Committee, Bowling Green State University.

² The authors wish to express their indebtedness to H. Lehtomaa, R. Knestrict, and the teachers of Dundee, Michigan, schools for their assistance in the collection of the data.

TABLE 1
NEUROTIC INDICES OF BOYS AND GIRLS HIGH AND LOW ON THE MASCULINITY SCALE

Group		Imp	Anx	BPI	Ho	Sch	Phy	Inf	Irr
High Boys	<i>M</i>	9.71**	16.83	22.36	3.18	1.40*	6.13	5.09	3.36
	<i>SD</i>	3.61	7.98	13.63	2.77	1.19	5.05	3.91	1.98
	<i>N</i>	48	48	45	45	45	45	45	45
Low Boys	<i>M</i>	8.26	17.48	24.81	3.76	1.88	6.88	5.53	3.77
	<i>SD</i>	3.60	8.21	12.56	2.33	1.55	5.28	3.38	2.09
	<i>N</i>	44	46	43	42	42	42	43	43
High Girls	<i>M</i>	8.09***	21.09*	25.17	4.71***	1.52	6.43	6.36	3.60
	<i>SD</i>	4.08	7.86	13.09	3.07	1.35	4.24	3.56	1.03
	<i>N</i>	44	45	42	42	42	42	42	42
Low Girls	<i>M</i>	5.92	18.56	22.45	3.04	1.43	6.27	5.57	3.43
	<i>SD</i>	2.66	7.32	10.83	2.41	1.26	4.52	3.34	2.02
	<i>N</i>	52	52	47	47	47	47	47	47

* Significant at the .10 level or less.

** Significant at the .05 level or less.

*** Significant at the .01 level or less.

RESULTS

The performances on the neurotic indices of boys and girls high and low on masculinity are presented in Table 1. The slight variation in the size of the *N*s is due to incomplete protocol on various tests. The results of Table 1 suggest that scoring high or low on the masculinity scale is reflected very little in independent measures of neurotic behavior for boys. The general tendency is for boys low on masculinity to appear more anxious

and neurotic, but the differences between them and the high scorers fail to achieve statistical significance. In the case of the girls, the results are more substantial. Girls high on masculinity tend to be more impulsive and anxious than girls low on masculinity, and show more neuroticism about their home life (parental rejection, etc.).

The performances on the neurotic indices of boys and girls high and low on femininity are presented in Table 2. The results indicate

TABLE 2
NEUROTIC INDICES OF BOYS AND GIRLS HIGH AND LOW ON THE FEMININITY SCALE

Group		Imp	Anx	BPI	Ho	Sch	Phy	Inf	Irr
High Boys	<i>M</i>	10.33**	18.26**	26.00**	3.87	1.71	7.03*	6.16**	3.61*
	<i>SD</i>	3.57	7.86	13.04	3.02	1.24	4.68	3.64	1.73
	<i>N</i>	43	43	38	38	38	38	38	38
Low Boys	<i>M</i>	8.52	14.31	20.73	3.85	1.52	5.27	4.41	2.85
	<i>SD</i>	3.60	6.67	9.99	2.29	1.52	3.87	2.68	1.97
	<i>N</i>	42	42	41	40	40	40	41	41
High Girls	<i>M</i>	7.49	19.86	25.25	4.20	1.80	6.65	4.80	3.05
	<i>SD</i>	3.32	7.13	13.41	3.19	1.54	4.58	6.16	2.10
	<i>N</i>	41	43	40	40	40	40	40	40
Low Girls	<i>M</i>	7.00	20.73	25.77	4.14	1.73	6.70	4.52	3.70
	<i>SD</i>	2.92	6.79	11.24	2.66	1.19	4.39	5.78	1.89
	<i>N</i>	48	48	44	44	44	44	44	44

* Significant at the .10 level, or less.

** Significant at the .05 level, or less.

TABLE 3

NEUROTIC INDICES OF BOYS AND GIRLS HIGH AND LOW ON THE MASCULINITY SCALE AND IN THE MIDDLE RANGE ON THE FEMININITY SCALE

Group		Imp	Anx	BPI	Ho	Sch	Phy	Inf	Irr
High Boys	<i>M</i>	9.02	16.90	24.30	3.60	1.50	6.15	5.50	3.60
	<i>SD</i>	2.96	8.31	11.36	2.62	1.29	4.52	3.40	1.88
	<i>N</i>	21	21	20	20	20	20	20	20
Low Boys	<i>M</i>	9.35	21.40	29.33	4.06	2.06	8.39	6.94	4.56
	<i>SD</i>	2.88	9.06	13.45	2.29	1.47	5.33	3.79	1.99
	<i>N</i>	17	18	18	18	18	18	18	18
High Girls	<i>M</i>	7.53	20.60	21.71	3.50	1.64	5.50	5.07	3.93
	<i>SD</i>	4.25	9.36	7.91	2.10	.78	3.31	1.76	1.87
	<i>N</i>	15	14	14	14	14	14	14	14
Low Girls	<i>M</i>	5.82	17.78	21.92	3.13	1.17	6.38	5.43	3.38
	<i>SD</i>	2.79	8.08	11.46	2.77	1.18	4.58	3.99	2.21
	<i>N</i>	27	27	24	24	24	24	24	24

that boys who are high on femininity are more impulsive, anxious, and neurotic than boys who are low on femininity, and that their neurotic concerns focus upon physical problems, feelings of insecurity, and irritability. From Table 2 it can be seen that girls high and low on femininity do not differ significantly in their performances on the neurotic indices.

In order to examine the possibility that it was not the variation in masculine or feminine identification but response-set which accounted for some of the differences obtained, a further analysis was undertaken. This analysis was similar to that described above, except that in order to decrease the possibility of response-set being of importance, boys and girls were chosen who were high and low on each scale, but were within the middle range on the other scale (i.e., their scores on the opposite-sex scale were between the 25 and 75 percentiles). Thus, one of the groups of boys and girls was of those who had high or low masculinity scores (upper and lower quartiles), but were within the middle range on femininity scores. Table 3 presents the results on the neurotic indices of boys and girls high and low on masculinity and middle range on femininity. Though none of the means achieve statistical significance, boys low on masculinity and middle on femininity tend to be more anxious and neurotic than

boys high on masculinity and middle on femininity, with the focus of concerns on the physical and inferiority subscales of the Brown Personality Inventory. For the girls, those high on masculinity and middle on femininity tend to be more impulsive and anxious, with concerns about physical symptoms.

Table 4 presents the results on the neurotic indices of boys and girls high and low on femininity and middle range on masculinity. There is only one instance in which significant differences are found, but boys high on femininity and middle range on masculinity tend to be more anxious and neurotic, with the focus of concerns about physical, inferiority, and irritability feelings. For girls, those low on femininity and middle on masculinity tend to score higher on measures of anxiety and neuroticism, with concerns surrounding the home. Apparently, both scales are of some use in the evaluation of sex role identification, though this confusion is reflected most clearly by the opposite sex scale. It is noteworthy that though they do not achieve statistical significance (possibly because of the decreased size of the *N*), a number of the differences between the highs and lows in this latter analysis are of some magnitude, and are in the same direction as the scores when response-set is not controlled.

DISCUSSION

The results of the present study show that the opposite-sex scale has a higher relationship to various measures of emotional stability than does the like-sex scale. This suggests that the opposite-sex scales are more effective in the discrimination of faulty sex role identification. The like-sex scales, though discriminating between high and low scorers of the same sex, were not equally effective in detecting faultiness of sex role identification as reflected in these independent measures of emotional stability. Apparently it would be most economical to use the opposite-sex scale in diagnosing sex role identification.

The measures of masculinity and femininity used in this study required children to indicate whether they played or did not play certain games. In addition, children were required to indicate whether they liked or disliked the games that they played. The fact that the responses of both boys and girls to play and game items of their own sex were not significantly related to measures of emotional stability suggests that the role expectations for each sex are fairly explicit in this area of like-sex behavior. Apparently, boys are clear about what boys are supposed to do as boys, and girls are clear about what girls are supposed to do as girls, so that irrespective of sex role confusion, both sexes may make enough conventionally expected

responses on the same sex scale for it not to be useful as an indicator of sex role confusion. On the other hand, the greater effectiveness of the opposite-sex scale suggests that role expectations of this sort may be more ambiguous. That is to say, children are less certain about how much interest they should show in the activities of the opposite sex.

Examination of the qualitative differences between boys and girls who are high scorers on the opposite-sex scale and the independent measures of emotional stability shows that there are important sex differences in the type of sex role confusion yielded by this scale. For example, the boys who are high scorers on the femininity scale appear to show considerably more confusion than the girls who are high scorers on the masculinity scale. In previous papers, the authors have shown that femininity in boys is associated with high scores both on anxiety and impulsiveness (Sutton-Smith & Rosenberg, in press-a, in press-b). Those findings are confirmed by the present study. In addition, the fact that the anxiety is expressed mainly on the subscales insecurity, irritability, and physical problems suggests that these boys have a very uncertain "self" picture. On the one hand they are anxious about their personal and physical selves; on the other, they are prepared to indulge in impulsive acting

TABLE 4
NEUROTIC INDICES OF BOYS AND GIRLS HIGH AND LOW ON THE FEMININITY SCALE AND IN THE MIDDLE RANGE ON THE MASCULINITY SCALE

Group		Imp	Anx	BPI	Ho	Sch	Phy	Inf	Irr
High Boys	<i>M</i>	10.13	18.27**	24.81	3.44	1.75	7.00	5.25	3.25
	<i>SD</i>	3.11	7.40	13.96	3.14	1.44	4.70	2.86	1.39
	<i>N</i>	15	15	16	16	16	16	16	16
Low Boys	<i>M</i>	9.50	13.27	20.52	4.19	1.29	5.19	4.19	2.57
	<i>SD</i>	3.42	7.10	9.01	2.72	1.16	3.28	2.58	1.87
	<i>N</i>	22	22	21	21	21	21	21	21
High Girls	<i>M</i>	7.20	19.67	25.00	3.62	1.74	6.79	6.79	3.58
	<i>SD</i>	1.91	6.70	12.38	3.22	1.58	3.66	3.87	2.03
	<i>N</i>	20	21	19	19	19	19	19	19
Low Girls	<i>M</i>	7.00	21.24	27.19	4.90	1.67	6.67	6.95	3.67
	<i>SD</i>	3.25	7.10	10.48	2.36	.98	4.59	3.19	1.67
	<i>N</i>	21	21	21	21	21	21	21	21

** Significant at the .05 level, or less.

out behavior ("I like to throw snowballs," "I like to chase fire engines"); and in addition, they show a greater than normal preference for feminine play activities and games. While this empirically derived picture is a paradoxical one insofar as impulsiveness and femininity would appear superficially to be contraries, it is similar to a finding of Hartley (1959) based on interview materials that there are some boys who choose such a feminine-impulsiveness as a defense. There is some agreement then that at least one important type of sex role confusion in children of this age group is to be found manifest in this feminine-impulsive syndrome. While the dynamics lying behind this syndrome are not at present apparent, the authors favor the view that impulsiveness is a means of warding off anxiety about sex role deviancy (abnormal identification with feminine role preferences) through pseudomale acting out behavior (Sutton-Smith & Rosenberg, in press-a).

The highly masculine girls do not show quite the same inconsistencies in their various scores. Their high scores on the impulsiveness scale (on which males have higher average scores than females) are consistent with their high scores on the masculinity scale. Not only are the girls' choice patterns consistent, they are also more culturally acceptable than the choice patterns of boys scoring high on femininity. It is not unusual for girls to choose masculine type activities in Western culture (Brown, 1958); in fact it is increasingly suitable for them to do so (Rosenberg & Sutton-Smith, 1960; Sutton-Smith & Rosenberg, in press-c). Nevertheless, the presence of high anxiety scores in these girls suggests that their masculine identifications do involve some conflict. Apparently, it is perceived as a conflict between themselves and their parents (higher scores on the home subscale) who are seen as severe and rejecting, rather than a conflict in their own self picture. Possibly, their masculine assertiveness may be a form of defense against parental rejection.

SUMMARY

The present study sought to examine the relative effectiveness of like-sex and opposite-sex scales in discriminating sex role identification. Each subject received scores on like as

well as opposite-sex scales, and three independent measures of emotional stability. Apparently, children scoring high or low on a like-sex scale do not differ significantly on measures of emotional stability. Children scoring high on opposite-sex scales tend to be more anxious, impulsive, and neurotic than children low on opposite-sex scales. Explanations of the differing symptoms of such sex deviant boys and girls are offered. From the results of the present study, there is some doubt that like-sex scales, as they are traditionally used, are as effective as heretofore suspected in discriminating sex role identification.

REFERENCES

- BROWN, D. G. Sex role development in a changing culture. *Psychol. Bull.*, 1958, 55, 232-242.
- BROWN, F. A psychoneurotic inventory for children between nine and fourteen years of age. *J. appl. Psychol.*, 1934, 18, 566-577.
- CASTANEDA, A., McCANDLESS, B. R., & PALERMO, D. S. The children's form of the manifest anxiety scale. *Child Developm.*, 1956, 27, 317-326.
- HARTLEY, RUTH. Sex role pressures and the socialization of the male child. *Psychol. Rep.*, 1959, 5, 457-468.
- HATHAWAY, S. R., & MCKINLEY, J. C. *Manual and booklet for the MMPI*. New York: Psychological Corp., 1943.
- HENRY, G. W. *Sex variants*. New York: Hoeber, 1948.
- ROSENBERG, B. G., & SUTTON-SMITH, B. The measurement of masculinity and femininity in children. *Child Developm.*, 1959, 30, 373-380.
- ROSENBERG, B. G., & SUTTON-SMITH, B. A revised conception of masculine-feminine differences in play activities. *J. genet. Psychol.*, 1960, 96, 165-170.
- ROSENBERG, B. G., SUTTON-SMITH, B., & MORGAN, E. E. Historical changes in the freedom with which children express themselves on personality inventories. *J. genet. Psychol.*, in press.
- STRONG, E. K. *Vocational interests of men and women*. Palo Alto: Stanford Univer. Press, 1943.
- SUTTON-SMITH, B., & ROSENBERG, B. G. A scale to identify impulsive behavior in children. *J. genet. Psychol.*, 1959, 95, 211-216.
- SUTTON-SMITH, B., & ROSENBERG, B. G. Impulsivity and sex preference. *J. genet. Psychol.*, in press. (a)
- SUTTON-SMITH, B., & ROSENBERG, B. G. Manifest anxiety and game preference in children. *Child Developm.*, in press. (b)
- SUTTON-SMITH, B., & ROSENBERG, B. G. Sixty years of historical change in the game preference of American children. *J. Amer. Folklore*, in press. (c)
- TERMAN, L. M., & MILES, C. C. *Sex and personality*. New York: McGraw-Hill, 1936.

(Received April 21, 1960)

TEST ANXIETY AND PERFORMANCE UNDER STRESS¹

ZANWIL SPERBER

Children's Hospital, Philadelphia

There has been much interest in the effects of anxiety on performance in recent years. In a series of studies Taylor, Spence, and others at Iowa showed that subjects (Ss) with a high anxiety level performed better on a variety of simple tasks than Ss low in anxiety (Spence & Farber, 1953; Spence & Taylor, 1951; Taylor, 1951; Wenar, 1954). With more complex tasks, anxiety level tended to have the opposite effect, low anxiety Ss performing better (Farber & Spence, 1953; Maltzman, Fox, & Morrisett, 1953; Montague, 1953; Ramond, 1953; Taylor & Rechtschaffen, 1959; Taylor & Spence, 1952). These results were explained in terms of Hull's theory of learning (Taylor, 1951; Taylor & Rechtschaffen, 1959; Taylor & Spence, 1952). Child (1954) criticized this theoretical approach for concentrating on the characteristics (simplicity or complexity) of the task while ignoring the category of responses Ss learn to make to the cues provided by their own anxiety.

Sarason and his colleagues have presented

¹ A report of part of a dissertation submitted to the University of Michigan in partial fulfillment of the requirements for the PhD degree. The writer wishes to thank E. Lowell Kelly, Chairman, and members of the doctoral committee for their aid and encouragement.

The cooperation of the United States Air Force was secured by Abraham Carp, Chief of the Motivation and Personality Measurement Division of the Human Resources Research Center (HRRC), now called the Air Force Personnel and Training Research Center. His interest in this research is greatly appreciated. The HRRC detachment at Sampson Air Force Base, commanded by Charles J. Meder, participated actively during the entire process of data collection. The assistance given by the entire group, and especially by Raymond J. Stubblebine and William McMackin, is gratefully acknowledged.

This research, however, was not sponsored by the United States Air Force, and the Air Force is not responsible for this report.

a more complex theory for explaining the role of anxiety in performance. Their theory predicts different effects from small and large amounts of anxiety, and takes into account the S's response to his own anxiety (Mandler & Sarason, 1952; Sarason, Mandler, & Craig-hill, 1952; Wiener, 1959). They assume that two kinds of anxiety responses may be aroused by a testing situation: those which are self-centered and ego-defensive, and those which are evoked and learned in the course of task performance and are directed toward task completion.

Small amounts of anxiety are held to improve performance by increasing the S's motivation and strengthening his task relevant responses. Larger amounts of anxiety lead to or strengthen the ego-defensive (R_A) responses. These responses are characteristic of the S rather than specific to the task, always available in the response repertoire and readily evoked. Since R_A responses are self-centered rather than task relevant they interfere with performance. Individuals with a high anxiety drive will have many R_A responses in their response repertoire. In contrast, in relation to the total number of responses available, low anxiety Ss will tend to make more task relevant anxiety responses. Results obtained with Yale students as Ss and using a variety of tasks (e.g., Koh's blocks, maze learning, digit symbol, motor learning, projective material) generally support the theory (Mandler & Sarason, 1952; Mandler & Sarason, 1953; Sarason & Mandler, 1952; Sarason et al., 1952; Wiener, 1959).

Sarason and Gordon (1953) also developed a measure of anxiety specific to the test stress situation to be studied, implicitly recognizing that anxiety need not function as a unitary drive, but might be a potential reaction of the individual which is only elicited under certain

conditions. Sperber (1959) has reported correlations between the scores obtained by Air Force recruits on Sarason's Test Anxiety Questionnaire and their scores on the Taylor Manifest Anxiety scale (1953) and Winne's Neuroticism scale (1951). Winne's interpretation of his results indicates that his scale can be considered a general measure of manifest anxiety (pp. 120-121). The highest correlation between the test anxiety and general anxiety measures was $+0.35$, suggesting that a general anxiety measure might not be an adequate gauge of an *S*'s proneness to anxiety in a test situation. Since the interest of the present research was in studying how anxiety, when evoked, operates to affect test performance, Sarason's measure of anxiety proneness specific to testing situations was used.

The purpose of the present study was to determine whether there is any difference in performance between: (a) high test anxiety *Ss* under high vs. low stress, (b) low test anxiety *Ss* under high vs. low stress, (c) high test anxiety *Ss* vs. low test anxiety *Ss* under high stress, and (d) high test anxiety *Ss* vs. low test anxiety *Ss* under low stress.

A further purpose was to bring additional evidence to bear on the different theoretical approaches to the problem of anxiety and performance presented by Taylor and Spence (1952) and by Sarason et al. (1952).

METHOD

Stress Environment

A situation is stressful by virtue of its ability to induce anxiety in individuals subjected to the situation. Lazarus, Deese, and Osler (1952) observed that "The principal problem in the study of behavior under stress has been the production of realistic stress situations" (p. 295). This experiment took place in association with a naturally stressful testing situation. The experimental data were collected at Sampson Air Force Base in December 1953, using new recruits as *Ss*. Prior to testing, an officer lectured them on the importance of the aptitude testing program they were to undergo. Many of the recruits hoped to benefit from the opportunity for some specialized vocational training which would equip them for a post service career; all were interested in getting a good assignment for their service period. The officer told them, in essence, that the realization of their hopes would be determined by the results of the aptitude testing.

The recruits were then tested on two successive days. The first day was devoted to the regular apti-

tude testing program. The second day was scheduled for HRRC experimental testing, but the recruits were only told to report to the same building for more testing. The use of identical administration procedures by the HRRC and aptitude testing staffs reinforced the *Ss*' tendency to view both days' testing as serving the same purpose.

High stress. *Ss* were allowed to continue believing that the experimental tests, presented as the Abilities Test Battery and administered by uniformed military personnel,² were part of the official assessment program. The battery consisted of a large number of timed subtests. The instructions implied that the *Ss* were expected to finish all of the items in each subtest, incompleting items counting as failures. The time limits were so chosen that the *Ss* could not possibly complete all of the items. The very first "test" was an interesting puzzle chosen to challenge and involve the *Ss*, but not solvable in the time allotted (Cowen, 1952, p. 514). It was included solely for the purpose of starting the *Ss* with a failure experience.

After completing the battery, a 40-minute interlude was devoted to procedures presented as not related to the "assessment program," and administered by a civilian. Then the sergeants announced: "We're going to repeat two of the tests you took this morning. You've had practice on them and you should improve. Let's see how much better you do this time." The first test was the puzzle, which again provided a failure experience. *Ss* then repeated the Letter Substitution Test.

Low stress. The instructions were designed to make it clear that the Abilities Test Battery was experimental, and not part of the regular assessment program, but also to indicate the importance of the tests to the Air Force and to maintain a high level of task oriented motivation for good performance. For these *Ss* the first test was a simple line tracing task which started them with a success experience. The rest of the battery was the same as for the high stress group. The low stress *Ss* were interrupted the same number of times and in the same places, but the instructions allowed them to treat unanswered items as evidence that the test required changes, rather than as signs of personal inadequacy.

After the 40-minute interlude, instead of receiving instructions which implied that an improvement in performance was expected, the low stress *Ss* were told: "We're going to repeat two of the tests you took this morning. You've had practice on them which should help, but you're somewhat tired by now—so let's see what happens. Do the best you can."

To help underscore the dissociation of the tests from the regular assessment program, the test administrator wore civilian clothes. To emphasize that the Air Force really had an interest in the procedure, the proctor, who was in a less active role, remained in uniform.

² Stubblebine administered the performance tests under both high and low stress conditions, and Swenson proctored all of the testing.

Performance Battery³

A series of paper and pencil tests, suitable for group administration, and tapping a range of cognitive functions was used. With the exception of the atmosphere reinforcing puzzle or tracing task, high and low stress Ss worked on the same tasks, which are described in the order they appeared in the battery.

Letter series. An expanded version of Thurstone's (1943) PMA Reasoning subtest, containing the original 30 items and 20 new ones constructed by the writer, was administered in two consecutive parts. For each part, containing a random selection of 15 original and 10 new items, 4 minutes were allowed, the parts being scored separately.

Letter substitution. The items were taken from Lazarus and Eriksen's (1952) expanded version of the Form I Digit Symbol test. The Ss were allowed 3½ minutes for 200 items. This test was repeated in the post interlude battery.

Revised Minnesota Paper Form Board. The series MB version of this test (Likert & Quasha, 1948), which measures the S's ability to perceive spatial relations, was administered in the usual manner.

Number matching. This test is similar to the number comparison sections of the Minnesota Vocational Test for Clerical Workers (Andrew, Patterson, & Longstaff, 1933), but new items were constructed by the writer. It measures the S's ability to concentrate and attend to detail, the items requiring the S to make rapid discriminations of small differences. Each page contained 100 items and was timed and scored separately. There were three consecutive parts to the test, with 4½ minutes allowed per page.

Test Reaction Questionnaire (TRQ)

After the performance testing, the experimental purpose of the tests was explained to the Ss. Then a 22-item questionnaire, constructed by the writer and similar in format to the Test Anxiety Questionnaire (Sarason & Mandler, 1952), was administered to ascertain the S's reactions to various aspects of the testing situation.⁴ The S put a mark on a 16-centimeter line to indicate the strength of his reaction, the ends of the line representing polar responses to the question. Each question was scored by measuring the distance in centimeters from the end of the line to the S's mark.

Personality measures

A slightly modified version of Sarason's Test Anxiety Questionnaire (TAQ)⁵ was administered to the

³ I wish to thank Robert H. Bauernfeind, Editor of the Test Department, Science Research Associates, for permission to use the Letter Series items from the PMA (Thurstone, 1943), and Richard S. Lazarus for making available the extended version of the Digit Symbol test.

⁴ Copies of the TRQ may be obtained from the writer.

⁵ I wish to thank Seymour B. Sarason for making available a copy of his scale, and granting permission

Ss after the TRQ. Following Sarason and Gordon's (1953) suggestion, local norms, based on the responses of both high and low stress Ss, were used in scoring each question. The resulting distribution of scores was very similar to that reported for Yale students (Sperber, 1959).

Other Procedures

The Minnesota Multiphasic Personality Inventory (MMPI) was administered in an afternoon session. Also available for each S was his stanine score on the Technician's Specialty Index (TSI) (Dailey, Lecznar, & Brokaw, 1948), considered to be the military test which best measures general intelligence level.⁶ For most Ss, raw scores on the Armed Forces Qualification Test (AFQT) (Boianovich, Mundy, Burke, & Falk, 1953), another gauge of intelligence level, were also available.

Subjects

All 399 new recruits who arrived at Sampson AFB to begin basic training the week before this experiment was conducted constituted the original sample tested. The Ss were organized in six "flights" with about 65 men in each. Three flights were arbitrarily assigned to each of the stress conditions. Tests were administered to one or two flights at a time, 201 men performing under high and 198 under low stress.

Recruits with a TSI in the lowest two stanines, or an AFQT score less than 15 were considered too low in intelligence to participate, and their data were not analyzed. Twenty percent were eliminated by these criteria. Ss whose MMPI records did not meet acceptable standards on the validity checks (Hathaway & Meehl, 1951) were also dropped, leaving 146 high stress and 148 low stress Ss. These men supplied the data on which were based the norms for scoring the TAQ, the correlations between test anxiety and general anxiety, and the analyses of reactions to the experimental conditions (TRQ responses).

The performance of only those Ss who scored in the highest and lowest quartiles on the TAQ was studied. High test anxiety (HTA) Ss had scores of 24 or higher, low test anxiety (LTA) Ss scored 11 or less. There were 61 Ss under high stress, 32 HTA and 29 LTA. Under low stress there were 71 Ss, 33 HTA and 38 LTA. These four groups were well matched with respect to age, years of education, and intelligence. For each group the mean age was close to 19, mean years of education was about 11, and the mean TSI stanine was approximately 5.5.

RESULTS

Effectiveness of the Experimental Operations

The distribution of responses of the 146 high stress and 148 low stress Ss to each of the questions of the TRQ were compared by

for such changes in wording as were necessary to adapt it for use in a different context.

⁶ Personal communication from Abraham Carp.

means of the Marshall test (Smith, 1953). The high stress Ss differed significantly from the low stress Ss in being convinced that the testing would decide their Air Force assignments. The low stress Ss more strongly accepted the idea that testing was for experimental purposes. They differed significantly ($p < .05$) on 11 of the 22 questions. High stress Ss felt more anxious, a greater pressure to do well, and a greater degree of unpleasantness in the testing situation. The two groups did not differ in their evaluation of how much their emotions and anxieties had interfered with performance.

The differences in TRQ responses of Ss scoring above and below the median on the test anxiety measure were tested separately for Ss who had performed under high stress and Ss who had performed under low stress. Under high stress, HTA and LTA Ss responded differently to 10 of the 22 questions (Marshall test, $p < .05$). HTA Ss more strongly believed that their Air Force assignment depended on their performance, felt more anxious, thought their emotions influenced their performance adversely, and considered the testing unpleasant. The groups did not differ with respect to test motivation.

Under low stress, HTA and LTA Ss responded differently to 4 out of 22 questions. There was a stronger tendency for HTA Ss to believe that their Air Force assignment might be at stake, and they responded with more anxiety to various aspects of the situation. There were no differences in motivation, feelings about the pleasantness of the situa-

TABLE 2
PERFORMANCE DIFFERENCES OF LOW TEST ANXIETY
SUBJECTS UNDER HIGH VS. LOW STRESS

Test	High Stress		Low Stress		<i>t</i> ^a
	Mean	σ	Mean	σ	
Letter Series 1	4.9	2.2	6.3	2.4	2.17*
Letter Series 2	6.2	2.8	7.9	3.0	2.42*
Letter Substitution A	98.0	19.0	108.9	29.2	1.84 ^b
Paper Form Board	39.6	8.5	39.8	6.3	.13 ^c
Number Matching 1	55.6	11.9	58.4	13.6	.90
Number Matching 2	55.9	10.4	57.0	12.8	.40
Number Matching 3	55.9	13.0	62.0	16.4	1.65
Letter Substitution B	133.5	23.3	137.6	33.2	.57 ^d
Intelligence (TSI)	5.2	1.5	5.7	1.6	1.34

^a $df = 65$.

^b Low Stress Ss were more variable ($F = 2.38$, $p < .02$); t and the value of t necessary for significance were calculated using the method described by Edwards (1950, p. 167-169).

^c High Stress Ss were more variable ($F = 1.83$, $p < .05$).

^d Low Stress Ss were more variable ($F = 2.03$, $p < .05$).

* $p < .05$.

tion, or evaluation of the influence their emotions had on performance.

Performance

The 32 HTA Ss tested under high stress performed significantly better than the 33 HTA Ss working under low stress on the three Number Matching tests. Critical ratios of the mean differences yielded p values of $< .01$, $< .001$, $< .05$. These results, however, were based on groups differing not only with respect to stress treatments, but also in test anxiety level. The high stress HTA Ss were significantly more anxious than the low stress HTA Ss (Marshall test, $p = .05$). To control for this the HTA groups were matched for level of test anxiety by a random selection procedure (Sperber, 1956, p. 136), 21 HTA Ss remaining in each stress group.

The results of the t tests of the differences in mean performance of the matched HTA groups working under high vs. low stress is given in Table 1. The groups also were adequately matched for intelligence. Stress clearly influenced the performance of HTA Ss, the groups tested under high stress performing significantly better on two tests, Number Matching 1 and 2.

The results of the t tests of the differences in mean performance of the LTA Ss tested under high vs. low stress are presented in Table 2. The groups were adequately matched both for intelligence and level of test anxiety. The LTA Ss consistently had higher mean scores under low stress. For both reasoning

TABLE 1

PERFORMANCE DIFFERENCES OF MATCHED HIGH TEST
ANXIETY SUBJECTS UNDER HIGH VS. LOW STRESS

Test	High Stress		Low Stress		<i>t</i> ^a
	Mean	σ	Mean	σ	
Letter Series 1	4.5	2.0	5.7	2.7	1.58
Letter Series 2	7.1	3.6	7.0	2.8	.09
Letter Substitution A	103.4	21.5	98.7	13.5	.83 ^b
Paper Form Board	39.6	8.8	40.6	8.0	.39
Number Matching 1	60.3	11.4	53.0	10.8	2.00*
Number Matching 2	61.4	10.2	52.9	10.9	2.54*
Number Matching 3	60.5	11.7	54.7	14.6	1.39
Letter Substitution B	133.5	27.9	129.3	17.9	.56 ^c
Intelligence (TSI)	5.4	1.5	6.0	1.5	1.22

^a $df = 40$.

^b High Stress Ss were more variable ($F = 2.53$, $p < .05$).

^c High Stress Ss were more variable ($F = 2.43$, $p < .05$).

* $p < .05$.

tests the mean differences reached statistical significance.

The *t* tests of differences in mean performance of HTA and LTA Ss, both tested under high stress are presented in Table 3. On six of the eight measures the HTA performed better. The only statistically significant difference was on Number Matching 2, where the HTA group was superior in performance. The results on the other two Number Matching tests supported this finding, both the mean differences favoring the HTA Ss, and approaching significance. For Number Matching 1 the value of *p* was .08, and for Number Matching 3 it was .10.

The critical ratios, testing the significance of the differences in mean level of performance under low stress of HTA and LTA Ss, are reported in Table 4. On all eight tests the LTA Ss were superior in performance. Two of the mean differences were statistically significant. For three more tests, namely, Number Matching 1 and 2, and Letter Substitution B, the differences were nearly significant, the values of *p* reaching .06, .09, and .07, respectively.

DISCUSSION

Our findings do not appear to be compatible either with the Iowa or with the Sarason theory. With respect to the Iowa position, we find that under high stress, HTA Ss perform better on Number Matching than do LTA Ss. Here higher drive seems to facilitate performance, the correct responses for this task presumably being uppermost in the average

TABLE 3

PERFORMANCE DIFFERENCES OF HIGH TEST ANXIETY AND LOW TEST ANXIETY SUBJECTS UNDER HIGH STRESS

Test	HTA		LTA		<i>p</i>
	Mean	<i>σ</i>	Mean	<i>σ</i>	
Letter Series 1	4.6	2.0	4.9	2.2	.43
Letter Series 2	7.2	3.7	6.2	2.8	1.11
Letter Substitution A	106.6	22.8	98.0	19.0	1.46
Paper Form Board	37.5	12.6	39.6	8.5	.72 ^b
Number Matching 1	61.0	11.0	55.6	11.9	1.81
Number Matching 2	62.1	10.2	55.9	10.4	2.31*
Number Matching 3	61.3	11.7	55.9	13.0	1.69
Letter Substitution B	134.8	28.9	133.5	23.3	.18
Intelligence (TSI)	5.4	1.5	5.2	1.5	.54

^a *df* = 59.

^b HTS Ss were more variable ($F = 2.19$, $p < .05$).

* $p < .05$.

TABLE 4

CRITICAL RATIO OF PERFORMANCE DIFFERENCES OF HIGH TEST ANXIETY AND LOW TEST ANXIETY SUBJECTS UNDER LOW STRESS

Test	HTA		LTA		<i>CR</i>
	Mean	<i>σ</i>	Mean	<i>σ</i>	
Letter Series 1	5.8	2.8	6.3	2.4	.81
Letter Series 2	7.1	3.1	7.9	3.0	1.15
Letter Substitution A	96.6	17.7	108.9	29.2	2.18*
Paper Form Board	39.2	7.3	39.8	6.3	.48
Number Matching 1	52.9	10.9	58.4	13.6	1.90
Number Matching 2	51.9	12.1	57.0	12.8	1.73
Number Matching 3	54.1	14.3	62.0	16.4	2.15*
Letter Substitution B	125.4	18.3	137.6	33.2	1.95 ^b
Intelligence (TSI)	5.5	1.6	5.7	1.6	.43

^a LTA Ss were more variable ($F = 2.73$, $p < .01$); *CR* necessary for significance at $p < .05$ with the observed heterogeneous variances is 2.03 (Edwards, 1950, p. 167-169).

^b LTA Ss were more variable ($F = 3.28$, $p < .001$).

* $p < .05$.

S's response hierarchy. On the same tests under low stress, however, LTA Ss perform better than HTA Ss, higher drive level apparently interfering with performance. Since the same test is involved in both cases, we can not readily attribute the reversal of the effects of a heightened anxiety drive to the different types of response hierarchies elicited by simple as contrasted with complex tasks.

Taylor (1956) has commented on the many characteristics other than drive level in which anxious and nonanxious Ss may differ, and which may influence performance (p. 303). The interest of the Iowa group in their study of anxiety has been restricted to the effects of drive, although the influence of anxiety drive per se on performance has been acknowledged to be small (Spence & Taylor, 1953; Taylor, 1956). Our own results confirm Taylor's suggestion that to understand the performance of anxious and nonanxious Ss, a broader approach to their characteristics is necessary. Such an approach will be indicated below.

The effects of anxiety on the performance of our Air Force recruits were consistently opposite to the results reported by Sarason. For the Yale Ss, strong anxiety—as manifested by Ss who were both high in test anxiety and tested under high stress—interfered with performance, but for the recruits studied in the present research strong anxiety improved performance. For the Yale students, moderate anxiety—as manifested by HTA Ss tested under low stress, or LTA Ss studied

under high stress—facilitated performance, but for our recruits it was associated with poorer performance. For the Yale students, a near lack of anxiety—such as would be the case for LTA Ss studied under low stress—was paralleled by a lowered level of performance. In contrast, for our Air Force recruits this lack of anxiety was associated with an improvement in performance.

A rationale for the performance of our recruits, and the difference between their performance pattern and that of Yale students can be offered in terms of the S's assessment of the personal importance of the testing situation, and the tendency to use avoidance or vigilance as a defense against anxiety, depending on the importance of a situation.

Douvan (1956) has demonstrated that an S's evaluation of the importance of a testing situation will depend on experiences associated with his social status. She showed that working class adolescents become involved and strive to do well in a situation where good performance can earn them an immediate reward, but are not motivated to do well when no direct reward is at stake. In contrast, the middle class adolescent appears to be strongly motivated for successful performance in both situations. Our Ss were predominantly from the working class (Sperber, 1959). McNeil (1953) has shown that a low evaluation of formal education is characteristic of families from a working class background. Although the recruits were about the same age as Yale students, a majority of them never completed high school. Their fewer years of education is assumed not to be simply a function of lower intelligence level, but also to reflect skepticism about the value of an academic education. This attitude was apparently shared by their parents since many of them were young enough to require parental permission to enlist. The Yale Ss were predominantly from the middle class (Sarason & Mandler, 1952) where characteristically there are strong pressures for academic achievement (McNeil, 1953).

The operation of vigilance and avoidance as defenses against anxiety is conceived in terms akin to Bruner and Postman's (1947) description of perceptual vigilance and defense:

In any given situation the organism singles out what it considers to be the environment's most relevant aspects—relevant to adaptation in the situation. So long as the situation is not *too* threatening or *too* exacting, avoidance of meaning may be emotionally the most economical response. But in situations which are highly threatening and highly exacting, the most adaptive perceptual response is frequently the one which takes most vigilant account of "reality" (p. 76).

With reference to the present findings, we view the high stress situation, testing which the S thought would determine his future, as one in which the S would be motivated to participate to the best of his ability. Participative rather than avoidance behavior seemed most in the S's self-interest, and we assume both HTA and LTA Ss were involved in the performance situation. We attribute the superior performance of the HTA Ss to the function of anxiety as an internal stimulus which could constantly remind them of both the importance and danger of the situation, and reinforce their motivation to be alert and perform well. The performance of the LTA group did not benefit from this increase in vigilance. This rationale seems especially plausible when we consider the nature of the Number Matching tests, on which the significant performance differences occurred. The greater anxiety of the HTA Ss would maintain their motivation to perform well on a repetitive, boring task, given at the end of a long session. An increment in perceptual vigilance would improve their discrimination of small differences.

Under low stress the LTA group performed better. The TRQ data allow us to assume that for the HTA Ss the situation was sufficiently like one involving real tests, to elicit feelings of anxiety. Since the instructions used assured the Ss that it was safe to be unconcerned about one's own performance, i.e., no immediate gain for them was possible, they would tend to defend themselves by withdrawal when the task became noxious. This need to avoid became strong enough to affect performance only on the more repetitious tasks used.

Our analysis of the results of the present study has been in terms which are similar in many respects to the theoretical formulations advanced by Lazarus and his colleagues (e.g., Vogel, Baker, & Lazarus, 1958) and by Rue-

bush (1960). The present study and other recent researches (Ruebush, 1960; Vogel et al., 1958; Wiener, 1959) have all presented results which point up the complex interaction of anxiety, motive, defense, and task variables. A prediction of whether anxiety leads to an improvement or decrement in the performance of a given *S* appears to depend on the answers to four related questions: (a) Does the *S* accept the performance situation as something so important that he must participate in it, or as something that can appropriately be avoided? (b) Does the anxiety aroused in the *S* by the situation lead to motivations which are more important as determinants of his behavior than the actual amount of anxiety elicited? (c) With respect to the direct effect of anxiety on performance, what kind(s) of defensive behavior is (are) engendered in the *S* by varying amounts of anxiety? (d) What relationship obtains between the nature of the *S*'s defenses and the structure and demands of the task?

SUMMARY

Air Force recruits, scoring in the highest and lowest quartiles on Sarason's measure of test anxiety were tested under high and low stress. The significant performance differences were:

1. Under high stress, HTA *S*s performed better than LTA *S*s on the Number Matching tests.
2. Under low stress, LTA *S*s performed better than HTA *S*s on the Letter Substitution and Number Matching tests.
3. Performance of HTA *S*s under high stress was superior to that of a group matched on test anxiety but subject to low stress on two of the three Number Matching tests.
4. Performance of LTA *S*s under low stress was superior to that of a group matched on test anxiety but subject to high stress on the two Letter Series tests.

The results were discussed in relation to the findings of Taylor and Spence, and those of Sarason et al., and a theoretical formulation advanced which emphasized the complex interaction of anxiety, motive, defense, and task variables.

REFERENCES

- ANDREW, D. M., PATTERSON, D. C., & LONGSTAFF, H. P. *Minnesota Clerical Test*. New York: Psychological Corp., 1933.
- BOLANOVICH, D., MUNDY, J. P., BURKE, L. K., & FALK, G. H. Development of the Armed Forces Qualification Test, Forms 3 and 4. *USA TAGO Personnel Res. Br. tech. res. Rep.*, 1953, No. 1078.
- BRUNER, J. S., & POSTMAN, L. Emotional selectivity in perception and reaction. *J. Pers.*, 1947, 16, 69-77.
- CHILD, I. L. Personality. *Annu. Rev. Psychol.*, 1954, 5, 149-170.
- COWEN, E. L. The influence of varying degrees of psychological stress on problem-solving rigidity. *J. abnorm. soc. Psychol.*, 1952, 47, 512-519.
- DAILEY, J. T., LECZNAK, W. B., & BROKAW, L. D. Development of the Airman Classification Test Battery. *Hum. Resources Res. Cent. res. Bull.*, 1948, No. 48-5.
- DOUVAN, ELIZABETH. Social status and success strivings. *J. abnorm. soc. Psychol.*, 1956, 52, 219-223.
- EDWARDS, A. L. *Experimental design in psychological research*. New York: Rinehart, 1950.
- FARBER, I. E., & SPENCE, K. W. Complex learning and conditioning as a function of anxiety. *J. exp. Psychol.*, 1953, 45, 120-125.
- HATHAWAY, S. R., & MEEHL, P. E. *An atlas for the clinical use of the MMPI*. Minneapolis: Univer. Minnesota Press, 1951.
- LAZARUS, R. S., DEESE, J., & OSLER, SONIA F. The effects of psychological stress upon performance. *Psychol. Bull.*, 1952, 49, 293-317.
- LAZARUS, R. S., & ERIKSEN, C. W. Effects of failure stress upon skilled performance. *J. exp. Psychol.*, 1952, 43, 100-105.
- LIKERT, R., & QUASHA, W. H. *The revised Minnesota Paper Form Board Test: Manual*. New York: Psychological Corp., 1948.
- MCNEIL, E. B. Conceptual and motoric expressiveness in two social classes. Unpublished doctoral dissertation, University of Michigan, 1953.
- MALTZMAN, I., FOX, J., & MORRISSETT, L. Some effects of manifest anxiety on mental set. *J. exp. Psychol.*, 1953, 46, 50-54.
- MANDLER, G., & SARASON, S. B. A study of anxiety and learning. *J. abnorm. soc. Psychol.*, 1952, 47, 166-173.
- MANDLER, G., & SARASON, S. B. The effect of prior experience and subjective failure on the evocation of test anxiety. *J. Pers.*, 1953, 21, 336-341.
- MONTAGUE, E. K. The role of anxiety in serial rote learning. *J. exp. Psychol.*, 1953, 45, 91-96.
- RAMOND, C. K. Anxiety and task as determiners of verbal performance. *J. exp. Psychol.*, 1953, 46, 120-124.
- RUEBUSH, B. K. Interfering and facilitating effects of test anxiety. *J. abnorm. soc. Psychol.*, 1960, 60, 205-212.
- SARASON, S. B., & GORDON, E. M. The Test Anxiety Questionnaire: Scoring norms. *J. abnorm. soc. Psychol.*, 1953, 48, 447-448.

- SARASON, S. B., & MANDLER, G. Some correlates of test anxiety. *J. abnorm. soc. Psychol.*, 1952, 47, 810-817.
- SARASON, S. B., MANDLER, G., & CRAIGHILL, P. G. The effect of differential instructions on anxiety and learning. *J. abnorm. soc. Psychol.*, 1952, 47, 561-565.
- SMITH, K. Distribution-free statistical methods and the concept of power efficiency. In L. Festinger & D. Katz (Eds.), *Research methods in the behavioral sciences*. New York: Dryden, 1953.
- SPENCE, K. W., & FARBER, I. E. Conditioning and extinction as a function of anxiety. *J. exp. Psychol.*, 1953, 45, 116-119.
- SPENCE, K. W., & TAYLOR, JANET A. Anxiety and strength of the UCS as determiners of the amount of eyelid conditioning. *J. exp. Psychol.*, 1951, 42, 183-188.
- SPENCE, K. W., & TAYLOR, JANET A. The relation of conditioned response strength to anxiety in normal, neurotic, and psychotic subjects. *J. exp. Psychol.*, 1953, 45, 265-272.
- SPERBER, Z. A study of the role of anxiety level and defense preference in performance under stress. Unpublished doctoral dissertation, University of Michigan, 1956.
- SPERBER, Z. The Test Anxiety Questionnaire: Scoring norms for a noncollege population. *J. abnorm. soc. Psychol.*, 1959, 58, 129-131.
- TAYLOR, JANET A. The relationship of anxiety to the conditioned eyelid response. *J. exp. Psychol.*, 1951, 41, 81-92.
- TAYLOR, JANET A. A personality scale of manifest anxiety. *J. abnorm. soc. Psychol.*, 1953, 48, 285-290.
- TAYLOR, JANET A. Drive theory and manifest anxiety. *Psychol. Bull.*, 1956, 53, 303-320.
- TAYLOR, JANET A., & RECHTSCHAFFEN, A. Manifest anxiety and reversed alphabet printing. *J. abnorm. soc. Psychol.*, 1959, 58, 221-224.
- TAYLOR, JANET A., & SPENCE, K. W. The relationship of anxiety level to performance in serial learning. *J. exp. Psychol.*, 1952, 44, 61-64.
- THURSTONE, L. L. *Primary Mental Abilities Test*. Chicago: Science Research Associates, 1943.
- VOGEL, W., BAKER, R. W., & LAZARUS, R. S. The role of motivation in psychological stress. *J. abnorm. soc. Psychol.*, 1958, 56, 105-112.
- WENAR, C. Reaction time as a function of manifest anxiety and stimulus intensity. *J. abnorm. soc. Psychol.*, 1954, 49, 335-340.
- WIENER, G. The interaction among anxiety, stress instructions, and difficulty. *J. consult. Psychol.*, 1959, 23, 324-328.
- WINNE, J. F. A scale of neuroticism: An adaptation of the Minnesota Multiphasic Personality Inventory. *J. clin. Psychol.*, 1951, 7, 117-122.

(Received April 25, 1960)

WECHSLER'S DETERIORATION RATIO IN CLINICAL PRACTICE

T. G. CROOKES

St. John's Hospital, Aylesbury, England

Studies on Wechsler's Deterioration Ratio have mostly suggested that it is of very limited value in diagnosing brain damage. None of those, for instance, quoted by Yates (1954) showed any very convincing differentiation of brain damaged from other groups. Even where it distinguished brain damaged from "normals," it did not show any appreciable difference between brain damaged and other psychiatric groups (Hall, 1952; Rogers, 1950), and it is this latter distinction which it is of practical value to make. The usual method is to take groups of known diagnoses, apply the tests to them, and compare the results. For reasons given below, this was not thought to be the most satisfactory way in this case, and it seemed to be of interest to examine the ratios in a group referred for routine purposes over a period of time, to look into the distribution of the values and compare them with the final diagnoses.

The actual ratio used is not exactly the same as Wechsler's. The same four "don't hold" tests are used, but only two "hold" tests: Vocabulary and Picture Completion. The score on these is doubled, and the percentage loss is calculated in the usual way with Wechsler's age allowances. Object Assembly was omitted because it does not seem to be a good hold test even on Wechsler's data (1944, p. 150), and this is a rather embarrassing test for adults, except the duller ones. Information was omitted to keep the balance of verbal and performance tests. The ratio will be referred to as DR.

PROCEDURE

Wechsler's test is applied routinely to almost all patients referred for psychological examination at the above hospital. It is a regional National Health Service hospital, catering for all kinds of mental illness.

The DR was calculated for all male inpatients referred over a period of 6 years, whatever the reason for referral, providing they had done all the tests necessary for the ratio. In all cases it was the Wechsler-Bellevue Scale, Form I, and where the test had been repeated, the figures of the first testing were used. The diagnoses were the final diagnoses made on discharge or death, and, where the patient is still here, the latest diagnosis at the time of writing supplied by the psychiatrist in charge of the case.

There are altogether 261 men, of whom 171 were less than 50 years of age at the time of testing, and 90 were 50 or over. The main comparison is between those in whom the diagnosis implies physical interference with the brain (the "organic" group) and the rest. The under-50 organic group consists of 23 patients, 12 epileptics, 3 head injuries, 3 toxic conditions (2 of them alcoholic), 2 cerebral syphilis, 1 confusional state in disseminated sclerosis, and 2 cases of unspecified brain disease. The over-50 organic group contains 22 patients, 12 with senile or presenile dementia, 1 alcoholic dementia, 5 toxic conditions, 2 cerebral syphilis, 1 head injury, and 1 cerebrovascular disease.

The remainder were divided into five broad diagnostic categories, to see if any type of score was peculiar to any group, and to see if the results agreed with the finding of Rogers (1950) that all maladjusted groups showed higher DRs than normals, but did not differ among themselves. The under-50 "non-organic" group consists of 30 depressives (including manics), 45 schizophrenics, 29 neurotics, 33 psychopaths, and 11 paranoid states (including paraphrenics). The over-50 nonorganic group contains 40 depressives, 1 schizophrenic, 10 neurotics, 9 psychopaths, and 8 paranoid states.

RESULTS

Table 1 shows the means and ranges of DRs for the organic and nonorganic groups.

In both age groups the organics have considerably higher values than the others. (Comparing organic and nonorganic means, for the under-50 groups, $t = 4.866$, $p < .001$; for the over-50, $t = 4.596$, $p < .001$.) But for practical purposes, one needs to know to what extent they can be differentiated by taking

TABLE 1
MEAN DETERIORATION RATIO, ORGANIC AND NONORGANIC

Group	Under-50			Over-50		
	N	M	Range	N	M	Range
Organic	23	+28.9	+59 to -22	22	+23.0	+57 to -18
Nonorganic	148	+ 8.8	+57 to -42	68	- 0.4	+42 to -52

some critical value or "cutting score." Table 2 gives the number falling above +20 (Wechsler's suggested critical score) and above +30, by age group, and whether organic or not.

If the two age groups are combined, we find using a score of 20 that 76.6% are correctly classified (organics 64.4% and nonorganics 79.2%). Using the score of 30, 85.1% are correctly classified (organics 53.3% and nonorganics 91.7%). The increase in total percentage correct obtained by raising the critical score from 20 to 30 is mainly due to the fact that the nonorganic group is much larger than the organic; by raising the score the discrimination of the nonorganic group is improved, that of the organic group is made worse, and as the former is so much bigger they contribute more to the total percentage. However, it will be noted that with the score of 30, the mean of the organic and nonorganic percentages (72.5) is slightly greater than their mean with the score of 20 (71.8). This mean can be considered as the percentage correct if the two groups are made equal in number; so the cutting score of 30 appears slightly better. In any case, it is probably of more value in this kind of problem to have a more extreme score, which nearly eliminates

one group, providing it leaves a substantial number of the other. For instance, a score which gave 100% nonorganics correct and 50% organics, would be of more practical value than one which gave 75% of each, although the total percentage correct is the same. In the first case you would be able to feel virtual certainty when some scores came up, but never in the second case. The whole general question of how proportions obtained in this way can be used in a practical situation, where the size of the populations from which you are drawing your samples is unknown, is considered in the discussion.

Table 3 shows the mean values of the DRs for the nonorganic patients divided into five diagnostic categories, again separated into the two age groups.

The most striking thing about Table 3 is that the younger patients consistently give higher values than the older ones, while within each group, there is little difference between diagnoses. This is fairly clear by inspection, but a simple analysis of variance was carried out for each age group (leaving out the schizophrenic in the over-50). In each case, the between diagnoses mean square was smaller than that within diagnoses. The schizophrenics and paranoid states, under-50, were combined (one can find reasons why they might be combined) and compared with the rest of their age group. This gives $t = 1.31$, $p > .1$. In the older age group, the neurotics were compared with the rest; $t = 1.63$, $p > .1$. It will be seen also that the high values of the DR are fairly evenly distributed among the various groups.

From the way in which the DR is calculated, the expected mean value for any age group is nought, if the group is comparable with Wechsler's standardization group. For the over-50 group, the mean is very close to

TABLE 2
PROPORTIONS DISTINGUISHED BY CUTTING
SCORES OF +20 AND +30

Group	Under-50		Over-50	
	DR < 21	DR > 20	DR < 21	DR > 20
Organic	7	16	9	13
Nonorganic	112 (75% correct)	36	59 (80% correct)	9
Group	DR < 31	DR > 30	DR < 31	DR > 30
Organic	10	13	11	11
Nonorganic	134 (86% correct)	14	64 (83% correct)	4

TABLE 3
DETERIORATION RATIO OF NONORGANIC PATIENTS BY DIAGNOSIS

Group	Under-50				Over-50			
	<i>N</i>	<i>M</i>	Range	<i>n</i> > 30	<i>N</i>	<i>M</i>	Range	<i>n</i> > 30
Depressives	30	+ 6.8	+47 to -42	3	40	+0.8	+42 to -52	3
Schizophrenics	45	+10.4	+57 to -34	5	1	—	+14	0
Paranoid states	11	+14.1	+31 to -14	1	8	+0.1	+26 to -36	0
Neurotics	29	+ 8.1	+46 to -26	4	10	-9.6	+32 to -29	1
Psychopaths	33	+ 7.2	+49 to -29	1	9	+2.6	+28 to -16	0
Total	148	+ 8.8	+57 to -42	14	68	-0.4	+42 to -52	4

nought, but not for the under-50 group; comparing their mean (+8.81) with its standard error (1.50) gives $t = 5.87$, $p < .001$.

DISCUSSION

In differentiating organics from nonorganics, the results here seem rather better than those obtained in most studies. One reason for this is probably that these are not selected groups of patients, but patients referred for information in the routine way. This means that in most cases the diagnosis was in doubt at the time of referral, and the deterioration was not marked. Gutman (1950), in summing up her study, says that hers were definite cases and that in less advanced cases even poorer differentiation would be expected. In general, the principle of taking extreme cases is excellent, but in this case it is not appropriate. The theory behind the ratio is that certain types of activity deteriorate more rapidly than others with aging and other forms of damage to the brain. It is not suggested that the other activities do not deteriorate at all. Clearly with gross dementia all abilities, including Vocabulary, disappear, cf. Yates (1956). It might be expected that the discrepancy would be most clear in the early stages, while later on all abilities declined to a common level.

Hall's (1952) organic cases are described as mostly moderately deteriorated, and he found only 5 out of 24 with ratios over +29 and only 7 over +19. His inclusion of Object Assembly may partly account for this. The group described by McFie and Piercy (1952), consisting of patients with localized cerebral

lesions, gave results fairly close to those of the present study. Of 56 DRs calculated, they had 24 (43%) over +29 and 37 (66%) over +19. For my whole organic group, the equivalent percentages are 53 and 67.

An interesting problem is the question of what weight can be given to an actual DR in a practical situation. If one has obtained a ratio greater than 30, what can be said about it? In this group, 18 out of 216 nonorganics obtained such a score, or 1 in 12; 24 out of 45 organics did so, rather more than half. So, in a sense, it can be said that such a score has a good probability of being organic; an organic case is much more likely to produce one than a nonorganic. However, if one considers the distribution of the 42 such scores obtained, only 24 are organic, and the probability of a given score of over 30 being organic is only 4/7, not much more than a half. In general, this is because the probability of an event's belonging to one of two categories depends not only on the relative frequency of the event within each category, but also on the relative frequency of the two categories in the population from which the event is drawn. This seems to be a general difficulty in all tests of this kind except where a certain score is found in virtually 100% of one category. The proportions found in one situation cannot be used for giving a probability in another situation where the constitution of the population is not known, as in most clinical situations.

The information has to be combined with data from other tests, and with qualitative indications on the Wechsler itself, for ex-

ample, differences on the Similarities (Hall, 1952) and "rotation" on the Block Designs (Shapiro, 1951), which are known to be related to the distinction being made. Another approach might be to attempt to eliminate individuals who come at the extreme of the distribution of DRs without having deteriorated. According to Wechsler (1944, p. 66), a DR of +20 is 2 Probable Errors from the mean, so that in a normal group one would expect to find about 9% above +20, and about 2% above +30. Subjects with the classical type of congenital reading disability, in which ordering and spatial organization are involved, might be expected to make poor scores on don't hold tests, especially Digit Span, Arithmetic, and Block Designs. There are three subjects in the present series with reading difficulties marked enough for it to be mentioned as one of their problems. They are all in the under-50 nonorganic group, and all have DRs greater than +29. It may be noted that the present group agrees with Wechsler's in the distribution at the other extreme of the scale. There are 8 minus ratios greater than 30 (3%) and 21 greater than 20 (8%).

Another thing associated with high DRs is low scoring on the test. Hold scores of 4 and don't hold scores of 2 do not suggest much absolute difference in the abilities concerned, but they would give a DR of +50. In the nonorganic under-50 group, there are 19 patients with IQ less than 80; 6 of these have DRs over +30, as opposed to 8 out of the other 129 with IQ over 79. This is not of much practical help, because low scores also occur in organic cases. In the organic under-50 group, there are 5 out of 23 with IQ below 80, and all 5 have DRs over +30. However, it does increase one's confidence in the significance of a high DR in a person of higher IQ.

As far as the under-50 nonorganic group is concerned, the study confirms the finding of Rogers (1950) that psychiatric groups show a higher DR than normals but do not differ among themselves. The over-50 nonorganic group, however, does not show more deterioration than the normal expectation. This curious age difference was also found, in psychiatric patients, by Garfield and Fey (1948). A similar age difference, in the same direction,

appears in the organic groups. One explanation could be that the don't hold tests are affected by either mental illness or aging, but the two factors do not have an additive effect when they occur together, so that the older patient is not affected more than is allowed for in the age corrections; alternatively, the illness, when combined with greater age, might affect the hold as well as don't hold tests. If patients were retested after recovery with the equivalent form of the test, on each hypothesis the younger ones would be the same on the hold tests, and better on the don't hold; on the first hypothesis the older patients would be the same on both types of test, on the second, they would be better on both.

The fact that mental illness has this tendency to raise the DR does not destroy the utility of the index in the diagnosis of brain damage, since the latter raises it significantly more. An index of over 30 gives a strong initial suggestion of brain damage, especially when the IQ is not low. Whatever method of assessing deterioration is used, an estimate of intelligence is needed in evaluating the results, so that Wechsler's Scale can be used to fulfil a double purpose.

SUMMARY

A modified form of Wechsler's Deterioration Ratio was calculated retrospectively for 261 male mental hospital inpatients who had been tested routinely over a period of 6 years. The scores were compared with the eventual diagnoses, and it was found that the ratio distinguished organic and nonorganic cases fairly well. In the nonorganic group, there was no variation corresponding to diagnosis, but there was confirmation of previous findings that psychiatric patients give larger ratios than normals. This, however, applied only to the younger group (under-50).

REFERENCES

- GARFIELD, S. L., & FEY, W. F. A comparison of the Wechsler-Bellevue and Shipley-Hartford scales as measures of mental impairment. *J. consult. Psychol.*, 1948, 12, 259-264.
- GUTMAN, BRIGETTE. The application of the Wechsler-Bellevue scale in the diagnosis of organic brain disorders. *J. clin. Psychol.*, 1950, 6, 195-198.

- HALL, K. R. L. Conceptual impairment in depressive and organic patients of the pre-senile age group. *J. ment. Sci.*, 1952, **98**, 256-264.
- McFIE, J., & PIERCY, M. F. Intellectual impairment with localized cerebral lesions. *Brain*, 1952, **75**, 292-311.
- ROGERS, L. S. A comparative evaluation of the Wechsler-Bellevue mental deterioration index for various adult groups. *J. clin. Psychol.*, 1950, **6**, 199-202.
- SHAPIRO, M. B. Experimental studies of a perceptual anomaly: I. Initial experiments. *J. ment. Sci.*, 1951, **97**, 90-110.
- WECHSLER, D. *The measurement of adult intelligence*. (3rd ed.) Baltimore: Williams & Wilkins, 1944.
- YATES, A. J. The validity of some psychological tests of brain damage. *Psychol. Bull.*, 1954, **51**, 359-379.
- YATES, A. J. The use of vocabulary in the measurement of intellectual deterioration: A review. *J. ment. Sci.*, 1956, **102**, 409-440.

(Received April 29, 1960)

VOCABULARY DEFICIT IN BRAIN OPERATED SCHIZOPHRENICS

ROY M. HAMLIN¹

University of Pittsburgh School of Medicine

AND ELAINE F. KINDER

Rockland State Hospital, Orangeburg, New York

Traditionally, vocabulary has been regarded as one of the functions affected least and latest by impairment due either to schizophrenia or to brain damage. In a recent study, Smith and Kinder (1959; Smith, 1958) report no significant loss in vocabulary for topectomized schizophrenics as compared with controls, 8 years after surgery. On a variety of other tests the brain operated subjects did show significant losses. Smith has suggested that significant losses tend to appear more frequently on measures involving sustained attention and perceptual organization than on measures involving vocabulary and verbal skills.

Williams, Lubin, and Giesecking (1959) report seemingly opposite results for non-schizophrenic subjects with a variety of brain conditions: vocabulary and verbal skills not only show clearly significant losses, but furthermore the amount of loss is as much, or even more, than on spatial tests. Subject samples, site and nature of tissue damage, lapse of time following the lesions, and the type of vocabulary test used raise many questions. The present study focuses on the last of these three factors: what is the effect of an oral test format, as compared with a multiple-choice format, on vocabulary scores following brain surgery?

¹ Now at the Veterans Administration Hospital, Danville, Illinois. The senior author's participation in this research was made possible by support from Western Psychiatric Institute, Henry W. Brosin, Director; as well as by support from the Topectomy Follow-Up Project with a research grant (M-1191) from the National Institute of Mental Health and funds and assistance from the Research Foundation for Mental Health, Inc. of the New York State Department of Mental Hygiene.

This factor of test format, and of the testing procedure, has received little attention in studies of deficit following brain damage. The Williams et al. study does comment on the test used. These authors base their conclusions on the Reading and Vocabulary (RV) Test from the Army Classification Battery (ACB). They offer evidence for the essential equivalence of the RV test and the Wechsler oral Vocabulary test, citing correlations of .76 between these two tests for Army recruits, for enlisted men, and for brain injured patients. They also note that the "majority of the items in RV require the S to read and understand a paragraph. Only a minority are standard vocabulary items."

Although the present study limits consideration to test procedure, some comment on the general and often contradictory evidence for vocabulary deficit following brain damage should be offered. Two of the most relevant and adequate studies are those of Yates (1956) and of Weinstein and Teuber (1957). Yates, on the basis of an extensive review, concludes that "vocabulary does decline in patients suffering from brain damage." This general conclusion may be supplemented by Weinstein and Teuber's careful consideration of differential effects when lesions involve specific brain areas. With pretest and posttest scores for both injured and control subjects, these latter authors report obvious loss in verbal and vocabulary scores some 10 years after focal lesions involving the left parietal-temporal areas, and/or with aphasia as a symptom. For other focal lesions, including frontal lobe lesions, they report little or no loss in vocabulary and similar tests.

In regard to the more immediate question of test format and test procedure, vocabulary deficit may well be in part a function of the vocabulary test used. In the previously cited study, Weinstein and Teuber (1957) found little verbal loss in certain brain injured groups; but these same subjects showed obvious loss on a hidden picture test. Is the multiple-choice test something like a hidden picture test, with the correct answer word "perceptually hidden" to some degree among three or four other words? Again, in the Smith and Kinder (1959) study, the Capps Homograph Test showed significant loss although oral vocabulary did not. The Homograph Test is a vocabulary test which introduces an additional element, requiring the subject to make a conceptual shift from one definition to another for the same word.

The present study used three test formats with subjects from the same sample studied by Smith and Kinder (1959): one oral, and two multiple-choice, tests of vocabulary. The first multiple-choice procedure *maximized* the factor of sustained attention. The second multiple-choice procedure *minimized* the factor of attention. The element of perceptual organization was assumed to be inherent in both multiple-choice procedures.

Since the subjects had taken Wechsler oral Vocabulary tests before surgery and again 8 years after surgery, the chief interest in the present study was in the new multiple-choice procedure which was designed to require both sustained attention and perceptual organization. The same words used in this multiple-choice procedure were then given orally and again under modified multiple-choice conditions to determine the comparability of these words to the Wechsler words, and to explore the extent to which each subject could achieve an approximate definition under any conditions. Three hypotheses were considered:

1. Topectomized schizophrenics will show significant loss on a multiple-choice test of vocabulary which *maximizes* the importance of sustained attention.

2. Topectomized schizophrenics will show no significant loss when the same words are then given as an oral vocabulary test. Smith and Kinder (1959) had already reported no

loss in oral vocabulary for these subjects 8 years after surgery. The present oral test was designed to insure that the specific word list used did not account for a difference in results.

3. Topectomized schizophrenics will show a tendency to loss when the same words are given a third time in a multiple-choice procedure *minimizing* the importance of attention. The number of correct definitions was expected to be greater than on the first multiple-choice procedure and less than on the oral procedure.

METHOD

Subjects. The subjects and surgical procedures have been described in detail elsewhere (Lewis, Landis, & King, 1956; Smith & Kinder, 1959). The 40 subjects included in the present study were all patients at Rockland State Hospital, Orangeburg, New York. Of these 21 were operated subjects, and 19 nonoperated controls. All were from the New York State Brain Research Project, had been diagnosed as schizophrenic, and had been included in the study reported by Smith and Kinder.

The operations had been performed 10 years prior to the present study. The subjects had been originally divided into a C group of older subjects (age range 47-58 as of 1949) and a D group of younger subjects (age range 21-38 in 1949). Surgery consisted of an orbital or a superior topectomy, both bilateral frontal lobe operations.

Thirty-eight subjects were assigned to 19 pairs of operated and control patients, matched for preoperative Wechsler Vocabulary. In 15 pairs, it was possible to match the subjects by chronological age group. In the other 4 pairs, they were matched for Vocabulary but not for age.

Tests. In addition to the Wechsler Vocabulary before and 8 years after surgery, the 40 subjects were given three new vocabulary tests 10 years after surgery. The five tests, with the designation used in Table 1, were as follows:

1. Pre: The preoperative vocabulary score is the average of two Wechsler oral Vocabulary raw scores obtained within a period of a few months before surgery in 1949. On occasion, only one Wechsler was given before surgery, rather than the usual two.

2. Post-8: The average of two Wechsler oral Vocabulary tests given within a 30-day interval 8 years after surgery.

3. Oral: The 60 words in the first civilian edition of the Army General Classification Test (AGCT, 1947) were given orally and scored as in the Wechsler.

4. MC-1: The first multiple-choice procedure consisted of the same 60 items from the AGCT. The stem phrase was presented on a card for 5 seconds: for example, "BIG means the same as . . ."

The stem card was then removed, and a 10-second delay followed. The four answer words from the AGCT were then presented, without the stem card, and the subject marked his response. The delay procedure emphasized the necessity for *sustained attention*.

5. MC-2: The second multiple-choice procedure employed the same materials. The four answer words from the AGCT were presented first, and the subject read them aloud. Mistakes in reading were corrected. The stem card was then presented and read aloud by the examiner, with the answer words still in view. The subject then gave his answer orally. At every step, the examiner encouraged maximum attention to each element in the task. This procedure minimized the effect of variable attention.

In the Oral, MC-1, and MC-2 procedures, a "base" and "ceiling" were used where obviously indicated; subjects responded to an average of some 45 of the 60 items. A correction for chance successes was used. The order of the tests was: MC-1, Oral, MC-2, with MC-2 given in a second session. The advantages and disadvantages of a rotated order for the tests were considered. With the number of subjects available and the anticipated variability of psychotic behaviors, it was felt that exactly the same procedure should be used with all subjects, in order not to lose the precision in analysis offered by pair matching. As previously indicated, the oral and second multiple-choice procedures were designed as control observations. The unexpected results with the second multiple-choice procedure will be discussed later.

Analysis. Results were analyzed by the sign test, and by t tests for differences between correlated means. p values refer to a one-tailed test.²

RESULTS

Table 1 presents the results for the five vocabulary tests in terms of the number of controls surpassing topectomized patients, for 19 pairs of subjects matched on preoperative Wechsler Vocabulary. Significant values for the sign test are indicated. The results will be presented under the following headings: MC-1, Oral, MC-2, Wechsler, and the three hypotheses.

MC-1. The first multiple-choice procedure incorporated both *attention* and *perceptual organization*, as well as word knowledge. It was expected that this procedure might differentiate the topectomized and control

² Ardie Lubin, and Elizabeth Engle of the Walter Reed Army Institute of Research have provided the authors with detailed statistical analysis of all the major comparisons reported. In all statistics reported, their results confirm the figures contained in this paper.

TABLE 1

NUMBER OF CONTROLS SURPASSING TOPECTOMIZED PATIENTS ON FIVE VOCABULARY SCORES, FOR NINETEEN PAIRS OF SUBJECTS MATCHED ON PRE-OPERATIVE WECHSLER VOCABULARY

	Pre	Post-8	MC-1	Oral	MC-2
Control Higher	8	14*	12	13*	16**
Ties	1	0	2	1	0

Note.—Pre = Preoperative Wechsler (1949).
Post-8 = Wechsler (1957).
MC-1 = Multiple-choice, with delay (1959).
Oral = New oral vocabulary (1959).
MC-2 = Multiple-choice, no delay (1959).

* $p = .05$.

** $p = .01$.

groups. It failed to do so. Neither the sign test (Table 1) nor the difference between means ($t = 1.41$) is significant. As with every vocabulary test given after surgery, the *consistent tendency* for the controls to surpass the operates is apparent. The mean superiority of the controls is 4.3 words.

Oral. Since Smith and Kinder (1959) reported no significant difference in oral vocabulary for these subjects, it was expected that the new Oral vocabulary test would not differentiate the topectomized and control groups. This expectation was further encouraged by the high correlation between the new Oral and Wechsler oral given 2 years earlier: for 41 subjects, this correlation was .89. For 19 pairs matched on preoperative vocabulary, 13 pairs show Oral scores higher for the control subject, 5 higher for the operate, and 1 tie. The sign test is significant at the .05 level, suggesting some loss in oral vocabulary due to brain surgery. The t for the difference between means, however, is 1.29 and not significant. The average superiority of the control over the operate is 2.3 words.

MC-2. The second multiple-choice procedure minimized the importance of sustained attention but retained the factor of perceptual organization. This vocabulary test format yielded a clearly significant difference associated with brain surgery. Of 19 controls, 16 surpassed the topectomized subject with whom they were paired. The sign test is significant at the .01 level. The t for the difference between means is 3.21, also significant at the .01 level. The mean superiority of the controls over the operates is 6.3 words—

nearly three times as much as for the same words given orally.³

The serial order in which the tests were given could be important: MC-1, Oral, MC-2. The control subjects showed a consistent increase in mean score on each later procedure: 30.9, 35.6, and 38.1. The topectomized subjects obtained the following mean scores on the respective tests in order of presentation: 26.6, 33.3, and 31.8.

Wechsler. Smith and Kinder, using analysis of covariance, report no significant loss in Wechsler oral Vocabulary 8 years after surgery for these subjects. Pair matching, however, indicates significant loss. For 19 pairs, the control surpasses the operate in 14 cases. The sign test is significant at the .05 level. The *t* for the difference between means is 2.46, also significant at the .05 level.

To confirm this evidence for loss in Wechsler vocabulary, the data on all 52 subjects used in the previous study were re-analyzed by pair matching. It was possible to match each of the 24 control subjects with a topectomized subject on preoperative Wechsler Vocabulary. Inspection indicates that the 4 remaining operated subjects obtained comparable scores. The sign test (with the control surpassing the operate in 18 pairs) and the difference between means ($t = 2.42$) are both significant at the .05 level. The sign test just misses significance at the .01 level ($p = .011$). The relative mean loss due to brain surgery is 2.6 words.

Hypotheses. Results for the three hypotheses follow:

1. The topectomized schizophrenics do not show a significant degree of relative loss on the multiple-choice vocabulary test requiring both sustained attention and perceptual organization. As with all the vocabulary tests given after surgery, the consistent tendency for the controls to surpass the operates is suggested.

³ The gain from Oral to MC-2, for the 19 pairs of subjects, was greater for the control 15 times, for the operate 3 times, with 1 tie. This manipulation of scores has questionable justification, but does suggest the effect of different test formats on the same words.

2. The topectomized subjects do show significant loss when the same words are then given orally. The sign test for the new Oral vocabulary indicates a difference in favor of the controls at the .05 level of confidence. The superiority of the controls need not be associated with the serial order in which the tests were given in the present study, since results for the Wechsler oral Vocabulary, given 2 years earlier, are comparable. Both the sign test and difference between means is significant at the .05 level.

3. The multiple-choice test which minimized the necessity for sustained attention clearly differentiates the operates from the controls. This test was given last in all cases. The topectomized schizophrenics show more loss, and more significant loss, in vocabulary on this test than on any of the others. The differential effect of "test procedure" is indicated, but both test format and order of presentation may contribute to the result.

DISCUSSION

The present study considered specifically the effect of test format on vocabulary scores after brain surgery. The results do suggest that test procedure may be an important factor. However, another factor in test procedure, the order of test presentation, may influence the results as much as the test format. This Discussion will consider briefly the general results of the study, and will then comment on the special question of test procedure.

The evidence for some loss in vocabulary in topectomized patients, as compared with controls, 10 years after surgery is not surprising. The vocabulary losses reported here are small in terms of the instruments' precision: the relative loss in Wechsler Vocabulary is some 2.5 words. The number of studies, employing comparable controls, with which these results can be compared is exceedingly small. Only a few reports present: control subjects acceptably matched with experimental subjects on the basis of tests before brain damage; comparable pretest and posttest scores, especially after the lapse of a number of years; and any reasonably adequate evidence for the location of the lesion. Perhaps the most nearly comparable study is

that of Weinstein and Teuber (1957) previously cited. These authors reported marked loss on verbal and vocabulary tests some 10 years after certain focal brain injuries, but little or no loss on such tests when only the frontal lobes were involved. The table which they present does indicate some tendency to *relative* loss for every injury group considered, whatever area of the brain was involved. That is, some injury groups gained over their preinjury scores, but the gain was never as great as the gain made by the control subjects. The present results are similar: slight, but in this case significant, loss in vocabulary 8 and 10 years after topectomies involving frontal lobe areas. Since Weinstein and Teuber (1957) report on a group test, on nonpsychotic subjects, and on injuries rather than bilateral operations, their study is only roughly comparable to the present one. The striking loss in verbal skills reported by Williams et al. (1959) must be evaluated in light of the heterogeneous brain conditions involved and the relatively short length of time following recovery. The evidence suggests that vocabulary deficit some 10 years after limited frontal lobe lesions is not great, but may be statistically significant under conditions such as those of the present study.

The effect of a specific test procedure employed in measuring vocabulary deficit has been given only incidental attention in previous studies. In one sense, the present results are clear. These topectomized subjects tend to show some deficit in vocabulary on every vocabulary test given 8 or 10 years after surgery. The effects of surgery are most clearly apparent, however, on the multiple-choice procedure MC-2, which was given last in order of testing. The effect of test procedure is indicated, but either test format or order of test presentation could account for the results.

In the present study, the effect of format and of order may be closely related. This suggestion calls for a final word on the special problem of demonstrating effects due to surgery when both experimental and control subjects are markedly psychotic. For such a purpose, the most suitable test may be one which minimizes the capricious and unpredictable vagaries of schizophrenic behavior,

and yet retains some element related to brain damage. The MC-2 procedure may have been most effective for this reason. Each element in the task was presented separately, and every effort was made to insure active attention. Reading errors were corrected, and the subject reviewed each correction. When the crucial answer was finally permitted, the schizophrenic patient's sporadic or "functional" inattention to reality may have been largely overcome. The resulting optimal capacity to define words and deal with perceptual organization could then constitute a measure which would demonstrate the effects of surgery. The importance of simplifying elements, insuring attention, and using repeated measures should be considered. Paradoxically, such precautions might result in a measure of some aspect of "attention" which would reflect the effects of brain surgery.

The most discriminative test, MC-2, was given *last* to all subjects, after they had had repeated exposure to the words involved. Of the two oral tests, the one with which the subjects had had previous experience discriminated best. Sheer's (cf. Lewis, Landis, & King, 1956) astute observations on "practice gains" are related to the importance of driving each element in the task home before expecting schizophrenic subjects to respond in a sufficiently comparable manner to permit measurement of differences. That is, *familiarity* with a test may serve to decrease unpredictable and irrelevant schizophrenic behaviors. Unless the idiosyncratic capriciousness of schizophrenic behavior is taken into account, relevant differences may be readily obscured.

SUMMARY

The study considered the effect of test procedure on measures of vocabulary deficit 8 and 10 years after bilateral topectomy. Three vocabulary tests, each employing the same words, were given to 21 operates and 19 controls. The three tests were a multiple-choice test maximizing the necessity for sustained attention, an oral test, and a multiple-choice test minimizing the importance of attention. The results and conclusions follow:

1. The topectomized schizophrenics showed a consistent tendency to vocabulary deficit on all tests, either oral or multiple-choice. The loss was statistically significant in most cases, but slight: for example, a mean loss of some two and a half words in Wechsler Vocabulary.

2. Vocabulary deficit was greater, and most significant, on the multiple-choice procedure which minimized attention, and which was given after the other two tests. This procedure differentiated the operates and controls at the .01 level of confidence. The mean superiority of the controls over the operates was nearly three times as many words as for the same words given orally.

3. The results suggest that the test procedure used may be an important factor in studies of vocabulary deficit. Relevant considerations include both the test format and the order of presentation. Both format and order may contribute to the results of the present study.

REFERENCES

- Army General Classification Test*. (First civilian edition) Chicago: Science Research Associates, 1947.
- LEWIS, N. D. C., LANDIS, C., & KING, H. E. *Studies in topectomy*. New York: Grune & Stratton, 1956.
- SMITH, A. Changes in psychological test performances of brain operated schizophrenics after an 8-year interval. Unpublished doctoral dissertation, Yeshiva University, 1958.
- SMITH, A., & KINDER, E. F. Changes in psychological test performances of brain-operated schizophrenics after 8 years. *Science*, 1959, 129, 149-150.
- WEINSTEIN, S., & TEUBER, H. L. Effects of penetrating brain injury on intelligence test scores. *Science*, 1957, 125, 1036-1037.
- WILLIAMS, H. L., LUBIN, A., & GIESEKING, C. F. Direct measurement of cognitive deficit in brain-injured patients. *J. consult. Psychol.*, 1959, 23, 300-305.
- YATES, A. J. The use of vocabulary in the measurement of intellectual deterioration: A review. *J. ment. Sci.*, 1956, 102, 409-440.

(Received May 9, 1960)

AN EVALUATION OF THE NORTHWESTERN INFANT INTELLIGENCE TEST, TEST B

BERNARD B. BRAEN

Onondaga County Child Guidance Center

The Northwestern Infant Intelligence Test, Test B (Gilliland, 1951) contains 40 items and is designed for use with infants between the ages of 13 and 36 weeks. Preliminary statistical work with the test (Gilliland, 1951, p. 16) suggests that it may be a more suitable instrument for the assessment of infant intelligence than other tests for this age period. This paper attempts to provide a systematic assessment of reliability and validity of the test as well as other quantitative and qualitative considerations.

METHOD

The subjects consisted of 100 adoptive or boarding home babies of both sexes between the ages of 13 and 36 weeks inclusive.¹ In the first phase of the study the Cattell Infant Intelligence Test (Cattell, 1947) was administered to the baby by a qualified examiner.² The second phase occurred 3 days later when the Northwestern Infant Intelligence Test, Test B was administered to the baby by a different examiner.³ The third phase involved the readministration of the Cattell by the original examiner⁴ when the baby was 18 months of age.⁵

¹ The Child and Family Service and the Department of Public Welfare, Children's Division, both of Syracuse, New York were the participating agencies in the study.

² Deep appreciation is extended to Marilyn Rothschild for her crucial part in this project.

³ The author administered all of the Northwestern tests.

⁴ After 45 retests the original examiner left the employ of the clinic. The author administered the remaining 19 tests.

⁵ The decision to retest the babies at 18 months with the Cattell was determined by two factors: (a) The correlation coefficient between the Cattell at 18 months and the Stanford-Binet (Form L) at 3 years was .67 (Cattell, 1947, p. 49). This coefficient suggests that the Cattell at 18 months appears to be measuring the same factor or factors as the Stanford-Binet at 3 years. (b) The results of this study were to have definite practical implications regarding the

A frequency distribution of age by week for the babies administered the Northwestern is shown in Table 1. The mean age was 22.56 weeks with a standard deviation of 6.71 weeks.

RESULTS

The mean IQ for the Northwestern was 91.42 with a standard deviation of 9.54. The mean IQ for the Cattell was 112.00 with a

TABLE 1
FREQUENCY DISTRIBUTION OF AGE BY WEEKS FOR
NORTHWESTERN-TEST B SUBJECTS
(*N* = 100)

Age in Weeks	<i>F</i>
13	8
14	6
15	4
16	6
17	5
18	4
19	5
20	9
21	2
22	3
23	4
24	6
25	4
26	1
27	9
28	2
29	2
30	4
31	3
32	2
33	3
34	3
35	5
36	0

adoption testing program at the Child Guidance Center. For this reason it was important to gather validating data with expediency but without sacrificing rigor.

TABLE 2
MEANS AND STANDARD DEVIATIONS FOR NORTHWESTERN AND CATTELL STANDARD
SCORE IQs FOR FOUR AGE GROUPS AND TOTAL GROUP

Age in Weeks	N	Northwestern		Cattell		Mean Difference	t
		M	SD	M	SD		
13-16	24	99.50	6.34	103.17	12.19	-3.67	-1.69
17-21	25	97.60	9.52	93.96	9.45	3.64	2.18*
22-27	27	103.89	8.35	102.59	8.87	1.30	.83
28-35	24	100.38	11.94	104.88	10.43	-4.50	-1.78
13-36	100	100.42	9.54	100.00	10.19	.42	.67

* Significant at the .05 level.

standard deviation of 15.21. On the basis of these findings it appears that the unequal means and standard deviations obtained with these tests make any direct comparison of IQs impossible. Further, without some sophistication with the concept of variability there may be a tendency to misinterpret the meaning of the IQ obtained with either test. In order to make comparisons between the two sets of scores, all the IQs on the Northwestern and Cattell were converted to standard scores with a mean of 100 and a standard deviation of 10. Also the 100 babies were divided into four age groups with approximately an equal number of subjects in each age group.

The converted means and standard deviations for the four age groups and the total group appear in Table 2.

From Table 2 it appears that variability of IQ from one test to the other occurs in all four age periods but the *t* tests reveal that it is most pronounced at the 13-16, 17-21, and 28-35 week levels. The *t* for the 28-35 week old group is artificially elevated due to the low ceiling on the Northwestern. By the IQ calculation method described by Gilliland (1951, p. 14) an infant of 36 weeks can only achieve an IQ of 112 on the Northwestern if he passes all 40 items. An infant of 35 weeks can only achieve an IQ of 115 if he passes all 40 items, and so on. Such a state of affairs could serve to lower artificially the mean IQ on the Northwestern for the 28-35 week old group, which would then contribute to a deceptive difference between the means of the Northwestern and Cattell.

The means and standard deviations of the difference between the Northwestern and Cat-

tell standard score IQs for the four age periods are reported in Table 3.

It appears from these data that the mean difference in standard scores for each group and total age period is no less than 6 and no more than 10 units. However, the standard deviations for each age period suggest that individual infants can vary from no difference in relative position on the two tests to a difference of as much as 25 standard score units from one test to the other. Since only 3 days intervened between the Northwestern and Cattell testings, these data indicate that at this age level, examiner, subject, and test reliability are difficult to achieve.

Reliability

The reliability of the Northwestern for the 100 babies 13-36 weeks of age was assessed through the odd-even method. The resulting coefficient corrected by the Spearman-Brown formula was $.95 \pm .01$.

Table 4 shows the odd-even reliability coefficients for the four age periods.

Gilliland (1951, p. 16) reported a corrected odd-even coefficient of .80 computed from

TABLE 3
MEANS AND STANDARD DEVIATIONS OF THE DIFFERENCES BETWEEN NORTHWESTERN AND CATTELL
STANDARD SCORE IQs FOR FOUR AGE PERIODS

Age in Weeks	N	Range	M	SD
13-16	24	1-25	8.88	6.70
17-21	25	0-22	7.00	5.59
22-27	27	0-22	6.26	5.49
28-35	24	0-22	9.38	6.44

TABLE 4

NORTHWESTERN ODD-EVEN RELIABILITY COEFFICIENTS FOR FOUR AGE PERIODS

Age in Weeks	N	r	SE
13-16	24	.64	$\pm .09$
17-21	25	.76	$\pm .05$
22-27	27	.88	$\pm .02$
28-35	24	.89	$\pm .02$

data obtained from 214 babies between 13 and 36 weeks of age from the Chicago area. Cattell (1947, p. 49) reports odd-even coefficients for her test for age periods comparable to the Northwestern age levels. At the very early age levels (13 weeks and 3 months) both the Northwestern and Cattell are less reliable than at the later levels, but in relation to the Cattell at the early age period the Northwestern appears the more reliable instrument.

In spite of the fact that the Northwestern permits a reliable representation of particular skills during the testing session, the obtained coefficients do not give information regarding consistency of performance over time. It would appear then that the reliability of the test for predictive purposes is limited.

The reliability coefficients for each of the four age groups may in themselves be unreliable because of the small number of subjects at each level. However, since the progression of coefficients followed closely those obtained with the Cattell it appeared that this factor was not significantly affecting the resulting coefficients.

Correlational Analysis

The correlation coefficients between the Northwestern, Test B and the Cattell for the four age groups and the total group are reported in Table 5.

These coefficients suggest that there is an element of communality in the tests and that this communal aspect is apparent for all ages. The magnitude of the correlation coefficients remains about the same for each age period even though the reliability of both tests increases with age. Such a finding suggests that

the tests have more in common at the 13-21 week period than the 22-36 week period.

The correlation coefficient of .58 between the Northwestern and Cattell for the whole age range may not reflect realistically the relationship between performance on the two tests because of the low ceiling on the Northwestern already described. In order to evaluate the effects of this artifact, the correlation was redone with the 28-35 week old group eliminated. The resulting coefficient was $.62 \pm .05$. It does not appear that the elimination of the oldest group markedly affects the extent of the relation between performance on the two tests.

Sex Differences

A critical ratio was computed between the means for 61 males and 39 females on the Northwestern. The resulting ratio of 1.55 indicates that the obtained differences between the means of the sexes could be accounted for by chance.

Validity

Validity was assessed by correlating performance of 64 subjects on both the Cattell and Northwestern at 13 to 36 weeks with their performance at 18 months of age on the Cattell. Table 6 shows the means and standard deviations for the Northwestern and Cattell IQs at 13 to 36 weeks and the Cattell IQs at 18 months.

The correlation coefficient between the Northwestern IQs at 13 to 36 weeks and the Cattell at 18 months was $.38 \pm .07$. The coefficient between the Cattell at 13 to 36 weeks and the Cattell at 18 months was $.39 \pm .07$.

Both coefficients are low enough to indicate

TABLE 5

CORRELATION COEFFICIENTS BETWEEN NORTHWESTERN AND CATTELL IQs FOR FOUR AGE GROUPS AND TOTAL GROUP

Age in Weeks	N	r	SE
13-16	24	.51	$\pm .10$
17-21	25	.63	$\pm .08$
22-27	27	.54	$\pm .09$
28-35	24	.42	$\pm .11$
13-36	100	.58	$\pm .04$

TABLE 6

MEANS AND STANDARD DEVIATIONS FOR NORTHWESTERN AND CATTELL IQS AT 13-36 WEEKS AND CATTELL IQS AT 18 MONTHS

Test	N	M	SD
Northwestern (13-36 weeks)	64	90.91	9.38
Cattell (13-36 weeks)	64	111.92	16.91
Cattell (18 months)	64	110.58	10.65

that little faith can be placed in predictive statements regarding intelligence for the 13 to 36 weeks age group. In fact only about 14% of the variance on the Cattell at 18 months can be accounted for by variation in the Northwestern and Cattell IQs at 13 to 36 weeks. At the 13 to 36 week period, factors such as rapid and varying growth rates, emphasis on sensorimotor skills, varying motivation, relatively subjective administration and scoring procedures, and examiner unreliability all seem pertinent sources of uncontrolled variance that serve to reduce the predictive power of both the Northwestern and Cattell.

SUMMARY

This study was designed to investigate the reliability, validity, and certain other features of the Northwestern Infant Intelligence Test, Test B. One hundred adoptive or boarding home babies between the ages of 13 and 36 weeks were tested with the Cattell. Three days later the Northwestern was administered. The Cattell was readministered to 64 of the 100 babies when they were 18 months of age.

The general findings were:

1. The IQs obtained on the Northwestern and Cattell were not directly comparable due to unequal means and standard deviations.
2. The Northwestern has a low ceiling at the upper age levels which prevents full expression of an infant's developmental skills between the ages of 30 and 36 weeks.
3. Odd-even reliability for the Northwestern and Cattell is similar for the 13 to 36 week age period but the reliability for the Northwestern is somewhat better than the Cattell at the 13-16 week period. The reliability of both tests improved with increasing age.
4. There appears to be a common factor in both tests at the 13 to 36 week period. This is especially true at the 13-21 week level—probably due to a preponderance of sensorimotor items on both tests for this age.
5. There was no significant sex difference found on the Northwestern.
6. The validity results indicate that prediction of a child's intelligence level at 18 months is a very risky procedure when based on performance on the Northwestern and Cattell at the 13-36 week period.

REFERENCES

- CATTELL, PSYCHE. *The measurement of intelligence of infants and young children*. New York: Psychological Corp., 1947.
- GILLILAND, A. R. *Northwestern Intelligence Tests: Test B, for infants 13 to 36 weeks old*. Boston: Houghton Mifflin, 1951.

(Received May 11, 1960)

JUDGMENTS OF ADJUSTMENT FROM TAT STORIES AS A FUNCTION OF EXPERIMENTALLY ALTERED SETS

BERNARD LUBIN¹

Indiana University Medical Center

Several recent investigations have demonstrated the susceptibility of the projective techniques to situational influences (Masling, 1960). The arousal of temporary motivational states (Mussen & Scodel, 1955), the relationship between the subject and the examiner (Bernstein, 1956), the subject's attitude toward the test (Feldman & Graley, 1954), and the manner in which the test is defined (Henry & Rotter, 1956), produce identifiable changes in projective test response. These studies usually end at the point of measuring changes in the protocols; the implication seems to be that the modified protocols would influence subsequent clinical judgment.

The clinical psychologist usually works in a setting in which practical decisions having great consequences for the individual and for the community must be made. Often, the referrer requests an opinion as to an individual's "adjustment." Since the question rarely specifies a criterion situation for the judgment and since the various objective measures of adjustment are fairly specific (Tindall, 1955), the judgment is frequently based on a global estimate of intellectual and/or "emotional" controls. Psychoanalytic formulations continue to influence much of clinical practice; thus emotional control often is taken to mean control of sexual and aggressive drives, and inferences about psychological status are based upon the degree and manner in which drive derivatives enter consciousness and behavior.

The purpose of this investigation is to inquire into some of the determinants of global judgments of adjustment. Specifically, the effect of TAT stories which have been elicited

under different instructions are judged for level of adjustment by clinical psychologists. Stories elicited under "Facilitating" instructions (prestige suggestion which placed high value on spontaneity and individuality) have already been shown to contain a higher amount of sexual and aggressive expression than stories elicited under "Neutral" or "Inhibiting" (prestige suggestion which placed high value on constraint and conformity) instructions (Lubin, 1960). The expectation is, therefore, that stories from the Facilitating condition will be judged as lower in adjustment than stories from either the Neutral or the Inhibited condition.

METHOD

In a previous investigation (Lubin, 1960), a random sample of 60 male college freshmen was tested in a 3×2 covariance design. The five conditions were: two Card conditions (two TAT cards with high pull of sexual content and two TAT cards with high pull of aggressive content), and three levels of Instructional Set (Inhibiting, Neutral, and Facilitating). Set was produced by means of prestige suggestion: the Facilitated group was told that "normal," "well-adjusted" people tend to let their imagination go as it is stimulated by the cards; the Inhibited group was told that normal, well-adjusted people are the master of their imagination and emotions; and the Neutral group was given innocuous instructions.² Subjects were randomly assigned to conditions, 10 subjects to a condition. Analyses indicated that Set produced a significant effect on sexual and on aggressive expression, and that the interaction between Set and Cards produced a significant effect on aggressive expression.

² A more detailed description of the methodology including the complete instructions can be found in the report of the previous investigation (Lubin, 1960).

¹ I wish to thank James Norton for assistance with problems of design and analysis.

Stimuli

The two experimentally treated TAT stories produced by each of the 60 subjects were stapled together. The stories were verbatim transcriptions and contained a notation as to reaction time. Within each envelope, the 120 stories (60 pairs) were divided into two sets of 30 pairs each, 60 stories told to TAT Cards 2 and 10, and 60 stories told to Cards 8BM and 20. The paired stories within each set were thoroughly scrambled.

Judges

Ten clinical psychologists, seven of whom possessed the PhD in clinical psychology and three who had a master's degree and at least 2 years of additional clinical experience, participated as judges.

Procedure

In the instructions, the judges were told that the stories were produced by 60 male subjects between the ages of 18 and 23. They were informed as to the TAT cards which had elicited the stories, and they were requested to rank the four major cues which they had used in rating the stories. Each judge received an envelope whose contents are described above. In addition, each envelope contained the following seven-point scale together with detailed instructions for its use.³

1. Well-integrated, happy person, socially effective
2. Only mild problems in essentially well-functioning person
3. Particular problems of some difficulty but social effectiveness maintained
4. Discomfort from problems severe enough to require therapy but ability to carry on
5. Acute neurotic problems but reality contact tenuously preserved
6. Severe neurotic problems with disorganization
7. Severe disturbance bordering on psychotic or prepsychotic

The seven-point scale was adapted from one developed and used by Dymond (1954) in an investigation of the effects of psychotherapy. Repeat scoring reliability of this scale was found to be .94.

The two stories of each subject were assigned one rating by each of the 10 judges. Analysis of variance was conducted on the 600 ratings of adjustment.

RESULTS AND DISCUSSION

It is important to note that the validity or accuracy of the ratings was not a subject of

³ A copy of the complete instructions has been deposited with the American Documentation Institute. Order Document No. 6631 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance \$1.25 for microfilm or \$1.25 for photocopies. Make checks payable to: Chief, Photoduplication Service, Library of Congress.

investigation. The stories had been produced by a sample of 60 male freshmen chosen randomly from the total population of male freshmen of a large university. It can be assumed that these subjects represented a sample of relatively well-functioning individuals. In order to preclude a skewed distribution of ratings, the judges were not told that the subjects were college students. However, neither were they told that the subjects represented a malfunctioning group, such as hospital or clinic patients. Under these conditions, one would expect that the effects of the stories themselves would be maximized.

The scatter plot revealed that the 600 ratings made by the 10 judges were distributed in a normal fashion over the seven scale points.

Table 1 presents the summary of the analysis of variance of the ratings of adjustment. It can be seen that the Instructional Set under which the TAT stories were elicited originally produced a significant effect on the judgments of adjustment. Neither Cards nor the interaction between Set and Cards was found to produce a significant effect on judgments of adjustment. Those aspects of the judgments determined by the sexual or aggressive stimulus patterns of the cards seemed to produce an effect of equivalent magnitude.

It seems reasonable to conclude, therefore, that the subject's attitude toward the testing situation can influence not only the test protocols (Lubin, 1960) but also may produce an effect on resulting clinical judgment.

Further information concerning the effect of Instructional Set upon judgments of adjustment is presented in Table 2. Since the scale was constructed so that higher scores represent judgments of poorer adjustment, the order of the means indicates that highest level of adjustment was rated for stories from the Inhibiting condition, lowest for stories from the Facilitating condition, with stories from the Neutral condition occupying the intermediate position. The Tukey Studentized Range Test reveals that the difference between the means of the Inhibiting and Facilitating conditions is large enough to be accepted with confidence, but that we cannot safely say where the Neutral condition falls in between the other two.

TABLE 1
ANALYSIS OF VARIANCE OF RATINGS OF ADJUSTMENT

Source	df	SS	MS	F
Between Sets	2	688.53	344.27	4.561*
Between Cards	1	7.34	7.34	.097
Set X Cards	2	30.41	15.21	.201
Subjects Within Set X Card Cells	54	4076.30	75.49	
Totals	59	4802.58		

* Significant at the .05 level.

Although the lists of story cues upon which the judges based their ratings were too varied to permit meaningful categorization and analysis, 8 of the 10 judges mentioned "amount of sex or aggression" as one of the four main cues which they used. The previous investigation (Lubin, 1960) demonstrated that Instructional Set produced a significant effect on both sex and aggressive expression and further analysis of the effect indicated that the effect was linear, i.e., highest expression of sex and aggression in the Facilitating condition, lowest in the Inhibiting condition, and an intermediate position for the Neutral condition. Thus there is a strong suggestion that judgments of adjustment based on TAT stories are significantly influenced by the amount of sex and aggression which is expressed.

These findings support the observations of Soskin (1954) that when the clinical psychologist is requested to make interpretations based on projective test protocols, his judgments tend to be biased in the direction of pathology. Also, Kenny and Bijou (1953) found that when clinical psychologists were asked to rank TAT stories according to their interpretive significance, they tended to give greater weight to contents which represented expressions of the sexual and aggressive drives.

The findings of this investigation point to some of the risks involved in "blind analysis." When the subject has a Set to be spontaneous and to individualize himself, he expresses more sex and aggression in his TAT stories (Lubin, 1960), and when the clinical psychologist, without knowledge of the subject's Set, makes a judgment of the subject's adjustment based on these same TAT stories his judgment tends to be biased in the direction of pathology to

the extent to which sex and aggression is expressed in the stories.

It might be objected that in practice the clinical psychologist does not make judgments of adjustment based solely on the TAT and certainly not on such a small number of cards as were used in this study. The first part of the objection must be granted; such an important judgment would be based upon a battery of tests rather than a single instrument. It should be noted, however, that within the test battery, the TAT is likely to be used in a variety of ways (Dana, 1956). In addition, interpretation of the TAT in the clinical setting is more often based on global judgment than on the few time consuming objective methods of analysis which have been proposed.

SUMMARY

In order to test the hypothesis that judgments of adjustment made by clinical psychologists are influenced by the Set under which a subject takes a projective test, two TAT stories from each of 60 subjects were rated on a seven-point scale of adjustment by 10 clinical psychologists. The stories previously had been elicited under differential Instructional Sets: Inhibiting ($N = 20$), Neutral ($N = 20$), and Facilitating ($N = 20$).

TABLE 2
MEAN RATINGS OF ADJUSTMENT
BY SET CONDITION

Set	Mean
Inhibiting	35.35
Neutral	38.35
Facilitating	43.55

It was found that Instructional Set significantly influenced the ratings: most pathology was rated for stories from the Facilitating condition and least for stories from the Inhibiting condition. Collateral data suggested that the amount of sexual and aggressive expression in the stories was the most important factor which influenced the ratings.

REFERENCES

- BERNSTEIN, L. The examiner as an inhibiting factor in clinical testing. *J. consult. Psychol.*, 1956, 20, 287-290.
- DANA, R. H. Selection of abbreviated TAT sets. *J. clin. Psychol.*, 1956, 12, 36-40.
- DYMOND, ROSALIND F. Adjustment changes over therapy from Thematic Apperception Test ratings. In C. R. Rogers & Rosalind F. Dymond (Eds.), *Psychotherapy and personality change*. Chicago: Univ. Chicago Press, 1954.
- FELDMAN, M. J., & GRALEY, J. The effects of an experimental set to simulate abnormality on group Rorschach performance. *J. proj. Tech.*, 1954, 18, 326-334.
- HENRY, EDITH M., & ROTTER, J. B. Situational influences on Rorschach responses. *J. consult. Psychol.*, 1956, 20, 457-462.
- KENNY, D. T., & BIJOU, S. W. Ambiguity of pictures and extent of personality factors in fantasy responses. *J. consult. Psychol.*, 1953, 17, 283-288.
- LUBIN, B. Some effects of set and stimulus properties on TAT stories. *J. proj. Tech.*, 1960, 24, 11-16.
- MASLING, J. The influence of situational and interpersonal variables in projective testing. *Psychol. Bull.*, 1960, 57, 65-85.
- MUSSEN, P. H., & SCODEL, A. The effects of sexual stimulation under varying conditions on TAT sexual responsiveness. *J. consult. Psychol.*, 1955, 19, 90.
- SOSKIN, W. F. Bias in postdiction from projective tests. *J. abnorm. soc. Psychol.*, 1954, 49, 69-74.
- TINDALL, R. H. Relationships among indices of adjustment status. *Educ. psychol. Measmt.*, 1955, 15, 152-162.

(Received May 12, 1960)

ORTHOPEDIC DISABILITY AS A FACTOR IN HUMAN-FIGURE PERCEPTION

AURELIA LEVI

Teachers College, Columbia University

There is perennial interest in the possible relationship between personality dynamics and various features of human-figure drawing. Though experimental studies frequently fail to support such a relationship, the impression continues that it does nevertheless exist, and an occasional study showing positive results helps to keep the impression alive.

Some of this ambiguity of result may be owing to the fact that even if it be true that a figure-drawing is a projection of its creator's body image (Machover, 1949), the projective process is necessarily limited by perceptual limitations, including perceptual prejudices (Postman, Bruner, & McGinnies, 1948). Indeed, this is the heart of the basic projective hypothesis. It follows then that before we can proceed to evaluate the peculiarities of a drawing—its omissions, overemphases, distortions, or whatever—in terms of its creator's dynamics, we would do well to establish the prejudicial influence of these dynamics on his perceptions. In other words, even assuming that a drawing is a projection, before we can attribute a bit of elaborate overemphasis in the drawing to its creator's undue preoccupation with the part in question, we have to insert a middle step and show that his perception—as yet uncomplicated by the act of bodying forth mental images through paper and pencil—has already been shaped by that particular preoccupation. This point has been made concisely by Silverstein and Robinson (1956): "The assumption that the physical body, the body image, and the drawn figure are in isomorphic relation" remains as yet unjustified. From their study they conclude that "this one-to-one relationship does not seem to exist" (p. 340).

It is only with the existence of prejudicial

influences on the perceptions of several diagnostic groups that the present study is concerned. The final step, relating these perceptions to the recreative process of figure-drawing, is beyond the scope of this study. Our task will be to separate Silverstein and Robinson's three isomorphs into two groups of two, and investigate a hypothesized one-to-one relation between the first two, the physical body and perceptions of the human figure (the body image). Our diagnostic groups differ from each other with respect to orthopedic disability—a comparatively simple, objective, distinguishing criterion.

This study hypothesizes a one-to-one relation between the physical disability and perceptions of the human figure, and that a particular orthopedic disability acts as a dislocating influence on perceptions of the relevant body part. Specifically, it is hypothesized that subjects with disability of the legs will be unusually sensitive to the legs in drawings of the human figure; and that subjects with disability of the arms will be unusually sensitive to the arms in drawings.

Since the group of back disabilities is a less homogeneous group than the other two, and has less well-defined etiologies which have been thought to be of psychic origin, it is further hypothesized that this group will show a greater resemblance to the characteristics of a control group of nondisabled.

PROCEDURE

Subjects

The experimental group was composed of 38 subjects with orthopedic disabilities as follows: (a) 12 with traumatic disability (fractures, amputations) to the arms, and no other disability; (b) 13 with traumatic disability to the legs, and no other disability; (c) 13 with a variety of low-back disabilities (spinal

TABLE 1

AGE, SEX, AND EDUCATION OF CONTROL AND DISABLED SUBJECTS

Subjects	N	Age		Sex		Education	
		\bar{X}	Range	M	F	\bar{X}	Range
Control	35	40.8	19-68	21	14	9.0	2-20
Disabled	38	41.9	16-67	28	10	8.5	0-20

fusion; laminectomy; arthritis of the lower back; uncomplicated, nonradiating low-back pain of undetermined origin), and no other disability. The control group was composed of 35 subjects without any orthopedic disability. As shown in Table 1, the age of the experimental group ranged between 16 and 67, mean age 41.9, and of the control group between 19 and 68, mean age 40.8. The number of years of education for the experimental group ranged from 0 to 20, mean 8.5 (median 9), and for the control group between 2 and 20, mean 9.0 (median 10).

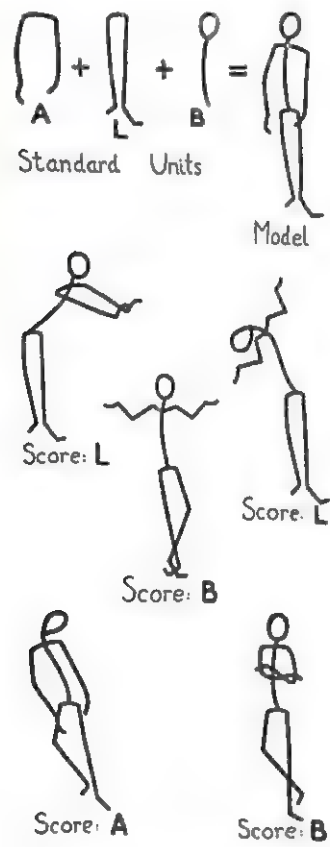


FIG. 1. Model figure, standard body units, and five examples of stick figures containing one standard unit each.

Instrument

The instrument was a set of 27 cards, each bearing a pair of stick figures (after Sarbin, 1954; Sarbin & Hardyck, 1955), and a single large card bearing a model stick figure. All the stick figures are composed of three parts: an arm unit, a leg unit, and a head-and-spine unit. The three parts making up the model are considered to be the standard units of reference; each of the 54 paired figures contains only one standard unit, the other two units being variants (Figure 1). Each stick figure appears twice in the set, once on the left and once on the right. The figures are so paired that 18 of them contain the standard arm unit, 18 contain the standard leg unit, and 18 contain the standard head-and-spine unit, but no 2 figures on one card contain any identical unit. Thus, from an objective point of view, each figure bears exactly the same amount of resemblance to the model as every other figure.

Method

Each subject was instructed as follows: "Here is a model figure, on the large card. I am going to show you other figures in pairs, and I want you to pick out which one of each pair looks more like the model, or resembles the model more closely." Scores for each subject are the number of choices he made involving each standard unit, the maximum number of choices possible for any unit being 18, and the minimum being 0. For example, a figure choice is credited to the Arm score if the standard unit it contains is the standard arm unit; similarly, it is credited to the Back score, if the standard unit it contains is the standard head-and-spine unit; etc. (Figure 1).

TABLE 2
COMPARISON OF MEANS OF THREE DISABILITY GROUPS WITH A CONTROL GROUP

Group	<i>N</i>	Mean	<i>t</i> ^a	<i>p</i>
A Choices				
Group A	12	14.17	3.49	.01
Control	35	8.94		
L Choices				
Group L	13	11.23	5.05	.001
Control	35	7.49		
B Choices				
Group B	13	12.92	1.48	.20
Control	35	10.54		

^a The *t* tests comparing Groups A and B with the control group were computed by a method designed for groups whose variances are unknown but presumed equal (see Table 3); the *t* test comparing Group L with the control group was computed for the case for which variances are unknown but presumed unequal (Walker & Lev, 1953).

TABLE 3

COMPARISON OF VARIANCES OF THREE DISABILITY GROUPS WITH A CONTROL GROUP AND WITH EACH OTHER

Group	N	Variance	F	p
A Choices				
Group A	12	20.7	1.02	ns
Control	35	20.29		
L Choices				
Group L	13	3.19	3.35	.05
Control	35	10.67		
B Choices				
Group B	13	18.41	1.47	ns
Control	35	27.02		
Group A	12	20.7	6.49	.01
Group L	13	3.19		
Group B	13	18.41	5.77	.01
Group L	13	3.19		
Group A	13	20.7	1.12	ns
Group B	12	18.41		

RESULTS

As was hypothesized, the arm-disabled and leg-disabled groups show a perceptual sensitivity that varies with the site of disability. For these two groups, means differ from those of nondisabled persons (Table 2), at a high degree of significance (.01 and .001, respectively).

As was hypothesized, the group of back disabilities shows no such perceptual difference from the nondisabled. The mean B Score for the back-disabled group does not differ significantly from the control mean (Table 2).

An unexpected finding is that the leg-disabled group is far more homogeneous than any of the others (Table 3).

DISCUSSION

The main finding of the study—that there is a one-to-one relationship between certain types of physical disability and perceptions of the human figure—provides a necessary, though not sufficient, step for a demonstration that a figure-drawing is a projection of its creator's body image.

That this one-to-one relationship is not a simple, across-the-board condition whose existence can be assumed for every kind of physical idiosyncrasy may be seen from the fact that the group of back disabilities, whose origins and symptomatology are much less clearcut than those of the other two groups, shows a less distinct perceptual prejudice; and from the differing variability of the groups.

It may be that the degree to which the disability is visually apparent immediately, exerts an important influence. Under such a scheme, low-back sufferers would be at the low end of a continuum of immediately apparent crippling; arm fractures and even amputations of the hand or arm—especially when fitted with a cosmetic prosthesis—while becoming apparent enough in any situation involving interaction, would nevertheless be more readily concealable than the impaired and distorted ambulation of the leg-disability group. Such an explanation might account for the findings with respect to the variability of the groups as well as their perceptual prejudice.

It should be noted that the results of this study may possibly be dependent on the fact that the control group and all three experimental groups were fairly homogeneous with respect to age, education, and socioeconomic level. To the extent to which differences in these factors might exert competing influences—as, for instance, the possibility that an extremely gifted, high *n* Achievement subject might bring into play an opposing perceptual prejudice akin to repression (Postman, Bruner, & McGinnies, 1948)—the pattern of results might be expected to vary; such an effect would, of course, be a matter for clarification through further research.

SUMMARY

This study compared three physical-disability groups with a group of nondisabled controls for their perceptual reactions to a structured test involving resemblances among schematized human-figure drawings.

It was ascertained that subjects with arm disability are particularly sensitive to the arms in a drawing, and that subjects with leg

disability are particularly sensitive to the legs in a drawing. Subjects with a variety of low-back ailments—often thought to be psychogenic—appear to be closer to the nondisabled control in their reactions than do either of the other two groups.

The group of leg disabilities is a much more homogeneous group than either the other two disabled groups or the controls.

A possible explanation that would account for all these findings is offered on the basis of the degree to which the disability is instantly apparent.

It is considered that this study gives support to a hypothesized one-to-one relationship between the physical body and the body image.

REFERENCES

- MACHOVER, KAREN. *Personality projection in the drawing of the human figure*. Springfield, Ill., Charles C Thomas, 1949.
- POSTMAN, L., BRUNER, G., & MCGINNIES, E. Personal values as selective factors in perception. *J. abnorm. soc. Psychol.*, 1948, 43, 142-154.
- SARBIN, T. R. Role theory. In G. Lindzey (Ed.), *Handbook of social psychology*. Vol. 1. Cambridge, Mass.: Addison-Wesley, 1954. Pp. 223-258.
- SARBIN, T. R., & HARDYCK, C. D. Conformance in role perception as a personality variable. *J. consult. Psychol.*, 1955, 19, 109-111.
- SILVERSTEIN, A. B., & ROBINSON, H. A. The representation of orthopedic disability in children's figure drawings. *J. consult. Psychol.*, 1956, 20, 333-341.
- WALKER, HELEN M., & LEV, J. *Statistical inference*. New York: Holt, 1953.

(Received May 18, 1960)

TAT PERFORMANCE AS A FUNCTION OF ANXIETY AND COPING-AVOIDING BEHAVIOR¹

E. JERRY PHARES

Kansas State University

A characteristic often associated with anxiety is its potential for generalization. Based perhaps on either qualitative similarity or identical elements, anxiety may become attached not only to the original arousal situation but also to other situations.

One such situation is that of projective testing. Anxious patients frequently seem to project threat into the test stimuli. For example, Rotter (1940, 1946) has stated that anxiety on the TAT is revealed in plots that emphasize sudden physical accidents and emotional trauma. Similarly, Schwartz (1955), from a Freudian viewpoint, has related the expression of castration anxiety on the TAT to themes in which occur genital or other body injury or loss, sexual or personal inadequacy, intrapsychic or extrapsychic threat, and loss of cathected objects.

However, to expect every anxious patient to produce such themes would seem a too simple approach to the problem. Not all anxious people handle their anxiety alike. Some subjects, having perceived a threatening stimulus, become evasive and produce bland stories, while others respond directly and create themes indicative of threat. The distinction drawn here parallels that of the perceptual defense-sensitization dimension (Carpenter, Weiner, & Carpenter, 1956).

In a similar vein, Weisskopf-Joelson, Asher, and Albrecht (1957) investigated "label-avoidance" as a manifestation of repression. They found some support for the hypothesis that people who strongly repress an impulse

tend to avoid expressing this impulse in situations carrying its label to a high degree. Applied to the TAT this could mean that subjects with strongly repressed sexual or aggressive impulses would frequently fail to give sexual or aggressive themes to pictures suggestive of an aggressive or sexual content.

The hypothesis of the present study is that high anxious subjects will show greater preference for TAT themes involving accident, threat, or trauma than will low anxious subjects when matched for tendency to evade or cope with threatening stimuli. The present study resembles one by Goodstein (1954) who found a nonsignificant relationship between anxiety and preference for anxiety-like TAT statements without controlling for coping-avoiding tendencies.

METHOD

Selection of High and Low Anxious Subjects

The Taylor anxiety scale (A scale) (1953) was used to select high and low anxious subjects. From 263 general psychology students who took a 50 item form of the A scale during pre-enrollment, 25 high anxious (scores above 18) and 25 low anxious females (scores under 7) were selected.

Determination of Coping-Avoiding Tendencies

This technique stems from a distinction between copers and avoiders recently made by Mainord (1956). He demonstrated that copers recalled more nonsense syllables associated with disturbing words than with neutral words, and avoiders recalled more syllables associated with neutral words. Goldstein (1959) utilized the same distinction in predicting differential responses to fear-arousing propaganda. Both of these investigations used an incomplete sentences blank (ISB) consisting of 40 critical and 20 filler stems. The former have direct sexual and aggressive implications. Critical items are scored on a three-point scale in terms of specificity, strength, and arbitrariness of response. A subject's score is the sum of

¹ Based on a paper delivered at the Annual Meeting of the Midwestern Psychological Association, St. Louis, Missouri, April 29, 1960.

This study was supported in part by a grant from the Faculty Research Fund of Kansas State University.

the weights assigned to each item and a high score indicates coping tendencies. Copers are thus subjects who respond most directly to the implications of the stems while avoiders are those who respond most evasively. Using this technique, two judges independently scored 13 protocols from a pretest population with 89% scoring agreement.

TAT Measure

To increase objectivity, the usual method of administering and scoring the TAT was modified. Seven cards (4, 6BM, 7BM, 13MF, 14, 17BM, and 20) were presented, each accompanied by six themes: four neutral and two embodying threat, accident, or trauma. For example, with 6BM:

neutral—This young man has come to the mansion to apply for the job of gardener. He has been asked to wait by this elderly maid. So, hat in hand, he waits.

threat—The son has just had to tell his mother that all of their savings have been wiped out. The stocks are worthless that they invested in. They are both stunned by this development. Nothing is left.

Each subject rank ordered the alternative themes in terms of how well they fitted the card. The subject's score is the sum of the ranks of the 14 threat themes. The higher the score, the lower the preference ranking for threat themes.

Each subject was tested individually. The modified TAT was administered, followed by the coping-avoiding ISB. Subjects responded anonymously to both TAT and ISB.

RESULTS AND DISCUSSION

From the foregoing procedure it was possible to match high and low anxiety subjects with respect to ISB scores. The N of each group was 19 and the mean ISB score for each group was 35.0.

The mean TAT score for the high anxious group was 47.2 (SD 4.7), and for the low anxious group 53.6 (SD 6.5). Applying a t test² the difference between the means is significant and in the predicted direction ($t = 3.4$; $p = .001$).

The results bear out the hypothesis that anxious people see more threat in TAT cards than do nonanxious people when their technique for defending against anxiety is controlled. Goldstein (1954) reported a nonsignificant trend in a similar direction due perhaps to a lack of control for the coping-avoiding dimension. Although a portion of his

data was based on 10 anxious and 10 non-anxious subjects while the present data is based on 19 pairs of subjects, the discrepancy between probability estimates in the two studies is much greater than would be expected merely by augmenting N .

Thus, it seems probable that in a given unselected population of anxious subjects, some who are copers and others avoiders, the confirmation of an otherwise perfectly logical and tenable hypothesis might be prevented. For example, Lesser (1959) demonstrated that under conditions of high anxiety about aggression, the intercorrelations among various measures of aggression were significantly lower than in the case of low anxiety over aggression. A behavior occurs not simply as a function of one variable but as a function of the relationship among several variables.

Generalization of the present results is, of course, limited by the sex composition of the population and by the modified TAT procedure.

SUMMARY

This study hypothesized that anxious subjects will show greater preference for TAT themes involving accident, threat, or trauma than will nonanxious subjects when matched for tendency to evade or cope with threatening stimuli.

Twenty-five subjects high on the Taylor anxiety scale and 25 low on that scale were administered a modified TAT and a special ISB which measures coping-avoiding tendencies. Subjects rank ordered neutral and threatening themes which accompanied seven TAT cards in terms of how well they fitted the cards.

With this procedure, 19 pairs of high and low anxiety female subjects matched for coping scores confirmed the hypothesis at a statistically significant level.

REFERENCES

- CARPENTER, B., WEINER, M., & CARPENTER, JANETH. Predictability of perceptual defense behavior. *J. abnorm. soc. Psychol.*, 1956, **52**, 380-383.
- GOLDSTEIN, M. J. The relationship between coping and avoiding behavior and response to fear-arousing propaganda. *J. abnorm. soc. Psychol.*, 1959, **58**, 247-252.

² A one-tailed test of the distribution of t was used.

- GOODSTEIN, L. D. Interrelationships among several measures of anxiety and hostility. *J. consult. Psychol.*, 1954, 18, 35-39.
- LESSER, G. S. Population differences in construct validity. *J. consult. Psychol.*, 1959, 23, 60-65.
- MAINORD, W. A. Experimental repression related to coping and avoidance behavior in the recall and relearning of nonsense syllables. Unpublished doctoral dissertation, University of Washington, 1956.
- ROTTER, J. B. Studies in the use and validity of the Thematic Apperception Test with mentally disordered patients: I. Method of analysis and clinical problems. *Charact. Pers.*, 1940, 9, 18-34.
- ROTTER, J. B. Thematic Apperception Tests: Suggestions for administration and interpretation. *J. Pers.*, 1946, 15, 70-92.
- SCHWARTZ, B. J. The measurement of castration anxiety and anxiety over loss of love. *J. Pers.*, 1955, 24, 204-219.
- TAYLOR, JANET A. A personality test of manifest anxiety. *J. abnorm. soc. Psychol.*, 1953, 48, 285-290.
- WEISSKOPF-JOELSON, EDITH, ASHER, E. J., & ALBRECHT, K. J. An experimental investigation of "label-avoidance" as a manifestation of repression. *J. proj. Tech.*, 1957, 21, 88-93.

(Received May 18, 1960)

RELATIONSHIPS BETWEEN DESCRIPTIVE CONTENT AND INTERACTION BEHAVIOR IN INTERVIEWS¹

JEANNE S. PHILLIPS, RUTH G. MATARAZZO, JOSEPH D. MATARAZZO,
GEORGE SASLOW

University of Oregon Medical School

AND FREDERICK H. KANFER

Purdue University

This study was undertaken in order to explore possible relationships between two different general approaches to the description and measurement of verbal interview behavior. One widely applied frame of reference directs attention to the communication aspects of verbal behavior, that is, to some symbolic dimension of the content of the words spoken, using content analysis to define and quantify its variables. A second frame of reference focuses upon quantitative *temporal* characteristics of interview interactions, utilizing measures such as number and duration of utterances, duration of silences, etc.

The few studies which have incorporated measures of *both* content *and* temporally defined variables have usually indicated the greater discriminatory power of the latter. For example, Page (1953), Lennard (1955), Goldman-Eisler (1952), and others have found that quantity or tempo of verbal output is both more stable and more highly correlated with various criteria of personal adjustment or psychotherapeutic success than are content-derived variables.

Lundy (1950) found that a single patient seen concurrently by two therapists differing in therapeutic technique showed no difference

in his responses to the two therapists in measures of content (Distress-Relief Quotient, Raimy's self-references, etc.). However, differences in *tempo* of interaction in the two sets of therapeutic interviews were apparent in the protocols. As a secondary aspect of a subsequent study, Lundy (1955) compared the temporal variables with clinical estimates of occurrence of significant emotional and topical content in "key" interviews. The results indicated that the contentual and temporal variables were related.

The present study involves the direct investigation of relationships between more precisely measured content variables and the temporal measures of the Interaction Chronograph. This instrument and Chapple's underlying interaction theory (Chapple, 1949; Chapple & Arensberg, 1940) constitute an extensive, systematically developed attempt to describe temporal phenomena in verbal behavior and to investigate their significance. Definitions of the Interaction Chronograph variables involved in the present study are given in Table 1. Without more extensive knowledge on our part of the subtle differences in "meaning" of the several inter-correlated Interaction Chronograph variables (Matarazzo, Saslow, & Hare, 1958), precise hypotheses as to specifically which content and interactional variables would be related could not be formulated. However, tentative hypotheses regarding the general types of relationships to be expected on a "face validity" basis were obviously the determinants of the kinds of content categories selected for use,

¹ This investigation was supported by a research grant (M-735) from the National Institute of Mental Health, of the National Institutes of Health, United States Public Health Service. This paper is based in part on a dissertation, under the direction of Frederick H. Kanfer, submitted by the senior author in partial fulfillment of the requirements for the degree of doctor of philosophy at Washington University. The data were collected while all of the investigators were at Washington University.

TABLE 1
INTERACTION CHRONOGRAPH VARIABLES

Pt.'s Units	The number of times the patient acted
Pt.'s Tempo	The average duration of each action plus its following inaction (silence), as a single measure
Pt.'s Silence	The average duration of the patient's silences
Pt.'s Adjustment	The durations of the patient's interruptions minus the durations of his latencies in responding, divided by Pt.'s Units
Pt.'s Initiative	The percentage of times, out of the available number of opportunities (usually 12) in Period 2, in which the patient acted again (within a 15-second limit) following his own last action
Pt.'s Quickness	The average length of time in Period 2 that the patient waited before taking the initiative following his own last action
Pt.'s Dominance	The number of times in Period 4 that the patient "talked down" the interviewer minus the number of times the interviewer talked down the patient, divided by the number of Pts.'s Units in that period

and are illustrated below in the description of the content system.

PROCEDURE

Forty patients randomly selected from new referrals to the Psychiatric Outpatient Clinic of a large urban medical center were interviewed by the same psychiatrist according to the published rules of the partially standardized interview which is used with the Interaction Chronograph in order that the interviewer may serve as a partially controlled or independent variable during each of five predefined subperiods of the interview (Matarazzo, Saslow, & Matarazzo, 1956; Saslow & Matarazzo, 1959). Each interview was observed from an adjoining room through a one-way mirror by the same observer, who recorded the ongoing interaction on an Interaction Chronograph and made a verbatim sound recording on a Gray Audograph. Ten of the 40 patients were omitted from the final experimental sample because of deficiencies in the sound recordings or, in a few cases, because of the presence of acute psychosis which resulted in confused, incoherent interview content. The 30 remaining patients served as the subjects of the present study. Of these, 17 were female, 13 were male. Their ages ranged from 20 to

61 years, with a median age of 35.5. The most frequent diagnoses were: hysteria (eight cases), anxiety neurosis (seven cases), depression (six cases) and schizoid personality (four cases).

The sound recordings were carefully and repeatedly monitored by the typist and a judge, until a high degree of recording transcription fidelity was achieved. Since the reliability of the content scoring process had been demonstrated on an independent sample of transcripts shortly before (Phillips, 1957), the final transcripts were unitized and then categorized, unit by unit, by one experimenter.

The content aspects of the verbal interview behavior were defined and quantified according to the category system schematically diagrammed in Table 2. The system and its development have been described in Phillips (1957). It represents an adaptation and extension of the Interpersonal System devised by Freedman, Leary, Ossorio, and Coffey (1951), C and C' in Table 2, which consists of a circular continuum of 16 categories, representing qualitative blendings of two orthogonal dimensions, love-hate and dominance-submission. A seventeenth category was added for coding of units unscorable or neutral within the Interpersonal System.

In addition to the Interpersonal System, several other dimensions were coded in order to achieve completeness of coverage and more general applicability. These other dimensions were operationally defined by: (a) coding of units without interpersonal reference, D and D' in Table 2, such as "I sat down and ate," "My head aches," etc.; (b) coding of the actor or subject of a description and of the receiver of action or attitude, if any, A and E in Table 2; and (c) coding of the general topic discussed, e.g., marriage, finances, symptoms, etc. Provision was also made for differentiation of descriptions of *motor acts*, C and D, from description of *nonmotor states* of being, thinking, feeling, etc., C' and D'. For example, "I yelled at her" is an interpersonal act, C, while "I was angry at her" is coded as an interpersonal state, C'; "I ate dinner" is a noninterpersonal act, D, while "I felt sick" is a noninterpersonal state, D'. Those units which referred directly to the ongoing interview interaction and had no referent outside of the current situation (e.g., "Thank you, doctor," "I agree with you") were classified separately, according to their function, within categories adapted from Bales (1950), II in Table 2. A residual category was utilized for unscorable units, III.

The resulting system is of a pyramidal nature, involving several series of mutually exclusive categories. It was considered particularly suited to the purposes of this study because of its interpersonal interactional emphasis and its use of content variables seemingly parallel to the temporal variables of the Interaction Chronograph. For example, it was expected that one or more of the Interaction Chronograph measures of verbal "output" (number and duration of utterances, etc.) would be related to the frequency of content describing the self as physically active, Self C+D in Table 2. Since the former measures of verbal output in the interview have been

TABLE 2
CONTENT CATEGORY SYSTEM

-
- I. Description (units which "tell about" an event)
- A. Actor (subject [person] of unit)
1. Patient himself
 2. Patient's family
 3. Patient's spouse, date, fiancée
 - etc.
- B. Time (of occurrence of described event)
1. Present
 2. Past
 3. Future
 4. Subjunctive
- C. Interpersonal Acts (overt motor behavior involving two or more people, e.g., telling, hitting, leaving someone)
- OR
- C'. Interpersonal States (nonovert states, thoughts, attitudes, etc., involving two or more people; e.g., thinking of, being angry at, or afraid of someone)
1. Interpersonal System Categories to 16
 17. Neutral, unscorable
- OR
- D. Noninterpersonal Acts (overt motor behavior involving only one person, e.g., eating, bathing, walking, etc.)
- OR
- D'. Noninterpersonal States (nonovert states, thoughts, attitudes, etc., involving only one person, e.g., feeling ill, being sleepy, poor, happy, etc.)
1. Positive (welcome, pleasant, etc. to patient)
 2. Neutral (neutral or indeterminant in significance to patient)
 3. Negative (disliked, unpleasant, etc. to patient)
- E. Object (person acted on or with, in interpersonal units)
(As for A above)
- F. Topical Area (of general life experience)
1. Financial
 2. Marital—Sexual
 3. Religious—Philosophical
 4. Educational
 - etc.
- II. Direct Interaction (units dealing directly with current interview interaction)
- A. Agrees, expresses compliance with interviewer
- B. Disagrees, expresses noncompliance with interviewer
- C. Asks for information, repetition, clarification, etc., from interviewer
- etc.
- III. Unscorable
-

described by Chapple² as involving both a physical high energy aspect and an out-going, interaction seeking aspect, it was further expected that one or more of them would be related to the degree of emphasis which the patient's content placed on interacting with others, Self C + C' in Table 2. Since the degree to which a patient takes the initiative in speaking, when the interviewer deliberately remains silent, has been proposed by Chapple as a measure of both the "drive" aspect of behavior and of the independence and scope of a person's interpersonal relationships, it was tentatively expected that Patient's (Pt.'s) Initiative would covary with content measures of breadth of interests (number of different persons and topics mentioned) and with description of the self in dominant interpersonal roles. It was such expectations as these, then, which guided the choice of dimensions of content to be included with the overly narrow Leary System in satisfying our goal of a comprehensive and multilevel content system.

The specific content scoring procedures include the stipulations that the categories are to be applied by the judges with a minimum degree of inference, and from the point of view of the patient. That is, all coding under this system is performed according to the relationship to or impact on the patient of the events as he describes them, without consideration of possible interpretations of psychological defenses or mechanisms, etc.

The content unit utilized for dividing verbalizations into countable and codable segments was basically defined as the minimal verbal statement which consensus of raters indicates to be understood as expressing an independent communication or thought. Although it was developed and tested independently, this unit is very similar to that of Auld and White (1956) and of Murray (1956).

Both the content unitizing and categorizing processes have been shown to have adequate reliability when applied independently by trained judges according to detailed definitions and rules (Phillips, 1957).

Raw frequency scores (number of content units coded for each category) were converted into percentages so that intersubject comparisons would be unaffected by differences in absolute numbers of unitized items. The percentage scores were transformed by the arc-sine transformation (Snedecor, 1946) for purposes of statistical analysis. The major content scores were then correlated (Pearson *r*) with 12 temporal Interaction Chronograph variables. Means and other statistics, when obtained, were converted back to percentages.

RESULTS

Table 3 presents the major findings. In order to conserve space, only those relationships which reached statistical significance, and those approaching significance (given in pa-

² E. D. Chapple. *Manual for the Interaction Chronograph*, personal communication, undated.

TABLE 3
CORRELATIONS BETWEEN CONTENT CATEGORIES AND INTERACTION CHRONOGRAPH VARIABLES

Content	Interaction Chronograph						
	Pt.'s Units	Pt.'s Tempo	Pt.'s Silence	Pt.'s Adj.	Pt.'s Init.	Pt.'s Quick.	Pt.'s Domin.
Description (I)			-.42	(-.34)		-.40	
Self Subject (A)				.48			.36
Self Acts (D + C)	-.38			-.40			
Self Interpersonal (C + C')	(-.34)						
Self Dominant-Hostile Quad.		.38		(.33)			
Self Submissive-Hostile Quad.							.37
Self Dominant-Positive Quad.				.41			
Self Submissive							(.33)
Self Positive				(.35)			
Self Noninterpersonal (D + D')	.37			-.36			
Self Noninterpersonal							
Negative (D + D' - 3)				-.38			
Other Interpersonal (C + C')			(-.33)	-.46			.36
Other Noninterpersonal (D + D')							
All Interpersonal (C + C')	-.40		.39	(.32)		.38	
Direct (II)	(.34)	(-.31)			.46	-.36	
Number Topics (F)	(-.31)		(-.31)	-.52	(.28)	(-.34)	
Number Persons (A + E)							

Note.— $r = .36$ for $p = .05$. $r = .46$ for $p = .01$. Parenthetical values approach significance at .05 level.

rentheses), are included. Table 3 also does not include content categories which are essentially reciprocals of those presented, nor some Interaction Chronograph variables which have been shown to be highly correlated with those included and whose correlations with content categories essentially duplicated these results.³

Pt.'s Units. Because of the relatively fixed length of the interview and of the interviewer's utterances, Pt.'s Units is highly correlated in a negative direction with measures (Pt.'s Action, Pt.'s Tempo, and Pt.'s Activity) of the duration of the patient's utterances. In the present sample it correlates $-.70$ with Pt.'s Action and may be assumed to be one of the more stable and representative measures of the general level of the patient's verbal output because, as a frequency measure, it is less affected by a few extremely short or long utterances than are the duration measures.⁴

³ The complete matrices for content vs. Interaction Chronograph correlations have been deposited with ADI. Order Document No. 6641 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance \$1.25 for microfilm or \$1.25 for photocopies. Make checks payable to: Chief, Photoduplication Service, Library of Congress.

The results shown in the first column of Table 3 reveal that the patient who has fewer Pt.'s Units, that is, who speaks relatively infrequently (but with longer durations per utterance), is shown by these correlations with content measures to describe himself as relatively: more active in his daily living, more oriented toward interactions with other people, less concerned with his own solitary experiences, interested in a wider variety of events in daily living, and less prone to evade description of himself in general.

Pt.'s Tempo. This is highly related to Pt.'s Units ($r = -.84$ in this sample) since the less frequently a patient speaks in the standardized interview, the longer of necessity will be his Tempo or "cycle" of speaking. Hence the relationships shown for Pt.'s Tempo should be considered in close conjunction with those for Pt.'s Units. As indicated in

⁴ The fact that these highly related measures of verbal activity level do, however, at times differ somewhat in their relationships with external behavior points out the necessity for retaining all of the Interaction Chronograph variables despite some high intercorrelations, until the areas of their overlap and independence have been more completely defined and explored.

Table 3, the longer the average duration of a patient's Tempo, the more he describes himself as dominant-hostile in his dealings with and attitudes toward other persons.

Pt.'s Silence. This is negatively related to the percentage of content units which are *descriptive* in nature and is positively related to the relative frequency of direct interaction with the interviewer. One might hypothesize that both silence behavior and remarks made to the interviewer represent "resistance" or avoidance tactics, since the questions asked of the patients are nondirective ones calling for description of their general life patterns and do not deal with events within the interview. This hypothesis is also supported by the tendency noted above for the number of (short) utterances to be inversely related to such direct discussion with the interviewer, as well as by the trend for longer silences to be accompanied by the introduction of fewer different topics.

Pt.'s Adjustment. Since all patients hesitated before responding for longer durations than they interrupted, in analyzing the data, the minus sign was omitted, so that high scores for Pt.'s Adjustment indicate greater relative latency of response. The results in Table 3 show that patients with high Pt.'s Adjustment scores can be considered to be almost opposite to those with fewer (but longer) Pt.'s Units. Thus, "maladjustment" is *negatively* related to the relative frequency of description of the self as active and to the number of different other persons mentioned, while it is positively related to the relative amount of talk about the self. Further, the "slow responders" relatively less frequently describe the acts or the attitudes, both interpersonal and noninterpersonal, of *other* people; when they do describe their own interpersonal attitudes or dealings with other people, they relatively more frequently characterize themselves as taking a submissive role, probably one which is also hostile. They also mention fewer negative noninterpersonal things about themselves while tending to focus more upon the noninterpersonal aspects of their lives in general. Similar to Pt.'s Silence, this latency-of-response measure tends to be negatively associated with amount of descriptive content and positively related to relative

frequency of direct interaction with the interviewer, supporting the hypothesis that this distinction between description and interaction with the interviewer represents avoidance in the face of difficulty in communicating.

The similar nature of the content category correlates of Pt.'s Units and Pt.'s Adjustment suggests what might be termed an "outward" or "other-directed" orientation in those patients who speak a fewer number of times and also in those who speak with a shorter latency of response. Patients who speak more times and those who wait longer before speaking have content which focuses more upon themselves, and thus might be termed "inward" or "self-directed." This interpretation is supported by our recent finding that schizophrenic patients, who might be thought of as being at the extreme pole on a continuum of inward-directed vs. outward-directed orientation, speak a significantly greater number of times (but in shorter average utterances), and with much longer average latencies before responding, compared to normals (Matarazzo & Saslow, 1961).

Pt.'s Initiative. The more the patient shows temporal interactional initiative (speaks again following his own last utterance), the more he also shows a form of "initiative" in raising new topics (and the broader one might therefore infer his interests or concerns to be). A similar but nonsignificant trend is shown for the total number of different persons mentioned.

Pt.'s Quickness. These results are supported by the finding that another temporal variable has similar correlations with the number of topics and number of persons mentioned. Unlike Pt.'s Initiative, however, Pt.'s Quickness is similar to Pt.'s Silence in being related significantly to the relative amount of descriptive content vs. interaction with the interviewer. Thus Quickness seems to have two components: one, the *readiness* to communicate in patients who take the initiative *rapidly*, a covariation which is similar to that found for Silence; and secondly, a component which, like the relationship found for Initiative, is related to the broader range of concerns (more people and more topics) of these same patients (with the shorter Quickness durations).

Pt.'s Dominance. This is directly related to the relative frequency with which the patient talks about people other than himself as well as to the relative concern he shows for the noninterpersonal feelings and behavior of others. Particularly striking is the finding that the more *dominant* the patient is in his temporal interview *behavior*, the more he describes himself in *content* as dominant-positive (e.g., teaching, helping, advising, protecting, etc.) in his attitudes and dealings with other people. Thus it seems that for the patient whose interpersonal style is described (in content) as more dominant, the interruption behavior on the part of the interviewer in Period 4 is animating and challenging, rather than defeating. The findings for Pt.'s Dominance are very similar to those for Pt.'s Units, the content correlates of which also emphasized relatively more discussion of others. However, patients with higher verbal output, as defined by Pt.'s Tempo, more frequently described themselves as dominant-*hostile* rather than dominant-*positive* in their own interpersonal roles.⁵

DISCUSSION AND SUMMARY

The results seem both encouraging in an exploratory study relating two quite disparate phenomena of interview behavior, and suggestive in the new meaningfulness which they add to the Interaction Chronograph variables by indicating relationships which seem inherently plausible and internally consistent. They provide a foundation for an approach to personality which combines content and temporal variables, and suggest personality dimensions which, underlying as they do at least two quite different spheres of behavior, may be particularly pervasive and consistent.

Viewed as a whole, the data presented in Table 3 suggest that patients who speak less often, who are faster to respond, and more dominant in the interview (that is, who have fewer and hence longer Units, shorter Pt.'s Adjustments, and more Dominance in the interruption stress period) have interview con-

tent which is relatively more oriented towards other people and towards interpersonal interaction, with social roles which are relatively more frequently described by them as dominant, either in a paternalistic or a hostile fashion. On the other hand, the correlations imply that the more a patient loses or submits to interruptions, is hesitant in speaking, and is less active verbally, the more his content emphasizes his own noninterpersonal concerns rather than interaction with others, and the more submissively hostile is his self-described role with other people.

In conjunction with the results of other validity-oriented studies of interview behavior correlates (Hare, Waxler, Saslow, & Matarazzo, 1960; Matarazzo, Matarazzo, Saslow, & Phillips, 1958; Matarazzo & Saslow, 1961), the present findings are a beginning at describing the characteristics which suggest a significant and cohesive description of the patient and how he interacts with others, viewed both from the subjective (content-inferred) and objective (temporally measured behavior) levels of observation. The major implication of these results bears upon the degree of generalizability of the Interaction Chronograph constructs and hence has to do with their concurrent (and, more remotely, construct) validity.

The correlations shown in Table 3 are small in the sense of accounting for relatively little of the variance, although respectable for complex personality variables. Further, the complete correlation matrix contained 336 *r*'s (28 content categories vs. 12 Interaction Chronograph variables), of which 28 were significant at the .05 level or better. Since 17 would have been expected to be significant at this level by chance alone, a replication study is being undertaken to determine whether the present findings can be cross-validated with another sample of subjects. A number of hypotheses are suggested by the results of this validity-oriented study which can be pursued in future investigations if the present findings are borne out in the replication study.

REFERENCES

- AULD, F., JR., & WHITE, ALICE M. Rules for dividing interviews into sentences. *J. Psychol.*, 1956, 42, 273-281.

⁵ There is no relationship in this sample between Pt.'s Units and Dominance ($r = -.27$, not significant), while Dominance and Pt.'s Adjustment are negatively correlated ($r = -.45$). Pt.'s Adjustment is positively correlated with Pt.'s Units, however ($r = .45$).

- BALES, R. F. *Interaction process analysis*. Cambridge: Addison-Wesley, 1950.
- CHAPPLE, E. D. The Interaction Chronograph: Its evolution and present application. *Personnel*, 1949, 25, 295-307.
- CHAPPLE, E. D., & ARENSBERG, C. M. Measuring human relations: An introduction to the study of the interaction of individuals. *Genet. psychol. Monogr.*, 1940, 22, 3-147.
- FREEDMAN, M. B., LEARY, T. F., OSSORIO, A. G., & COFFEY, H. S. The interpersonal dimension of personality. *J. Pers.*, 1951, 20, 143-161.
- GOLDMAN-EISLER, FRIEDA. Individual differences between interviewers and their effect on interviewees' conversational behavior. *J. ment. Sci.*, 1952, 98, 660-670.
- HARE, A. P., WAXLER, NANCY, SASLOW, G., & MATARAZZO, J. D. Simultaneous recordings of Bales and Chapple interaction measures during initial psychiatric interviews. *J. consult. Psychol.*, 1960, 24, 193.
- LENNARD, H. L. Concepts of interaction. Unpublished doctoral dissertation, Columbia University, 1955.
- LUNDY, B. W. An investigation of the process of psychotherapy. Unpublished master's thesis, University of Chicago, 1950.
- LUNDY, B. W. Temporal factors of interaction in psychotherapy. Unpublished doctoral dissertation, University of Chicago, 1955.
- MATARAZZO, J. D., & SASLOW, G. Differences in interview interaction behavior among normal and deviant groups. In I. A. Berg & B. M. Bass (Eds.), *Conformity and deviation*. New York: Harper, 1961, in press.
- MATARAZZO, J. D., SASLOW, G., & HARE, A. P. Factor analysis of interview interaction behavior. *J. consult. Psychol.*, 1958, 22, 419-429.
- MATARAZZO, J. D., SASLOW, G., & MATARAZZO, RUTH G. The Interaction Chronograph as an instrument for objective measurement of interaction patterns during interviews. *J. Psychol.*, 1956, 41, 347-367.
- MATARAZZO, RUTH G., MATARAZZO, J. D., SASLOW, G., & PHILLIPS, JEANNE S. Psychological test and organismic correlates of interview interaction patterns. *J. abnorm. soc. Psychol.*, 1958, 56, 329-338.
- MURRAY, E. J. A content-analysis method for studying psychotherapy. *Psychol. Monogr.*, 1956, 70(13, Whole No. 420).
- PAGE, H. A. An assessment of the predictive value of certain language measures in psychotherapeutic counseling. In W. V. Snyder (Ed.), *Group report of a program of research in psychotherapy*. University Park, Pa.: Pennsylvania State University, 1953.
- PHILLIPS, JEANNE S. The relationship between two measures of interview behavior comparing verbal content and verbal temporal patterns of interaction. Unpublished doctoral dissertation, Washington University, 1957.
- SASLOW, G., & MATARAZZO, J. D. A technique for studying changes in interview behavior. In E. A. Rubinstein & M. B. Parloff (Eds.), *Research in psychotherapy*. Washington, D. C.: American Psychological Association, 1959.
- SNEDECOR, G. W. *Statistical methods*. (4th ed.) Ames: Iowa State College, 1946.

(Received May 27, 1960)

THE CONCEPT OF NORMALITY:

A REPLY TO FREIDES

MAURICE KORMAN

University of Texas Southwestern Medical School

In an interesting paper published in this journal, Freides (1960) presses for the elimination of the concept of normality on the grounds that there is little agreement concerning its definition, that it ascribes absolute, yet culture bound, patterns of behavior and that it disregards the flexible interactions between personality and circumstances. Although Freides concerns himself nearly exclusively with the idealist-adjustment view of normality, other approaches such as the statistical-average conception are likewise dismissed with the conclusion that "for purposes of scientific theory and also for practical clinical purposes" our emphasis must shift away from considerations of normality (or pathogenicity) and toward a greater concern with "the potentialities of every person under the proper conditions." Such a position, it will be argued here, springs from too narrow a conceptualization of both scientific theory and clinical purposes. It has, moreover, some unfortunate implications for future research developments and basic orientations in clinical psychology.

Although normality is (today) an admittedly low powered concept, judgments of normality-pathogenicity hover in the background of many diagnostic and therapeutic decisions. It seems that, however ambiguous and ill-defined a conception of normality our professional behavior reflects, this term nonetheless behaves the way a good construct should (Beck, 1953): as a *summarizer* of a host of personality characteristics and behavioral tendencies, as a gross *predictor* of a person's future behavior and as an *object of search*. This last role will be made clearer as this discussion develops.

If our clinical practice reflects (as I believe

it does) the use of *some* concept of normality, it becomes appropriate to inquire into what kind of meaning can be ascribed to it. The controversy over the idealist-adjustive vs. the statistical-average interpretations of normality has obscured some more basic considerations. Why should the question, "What is normality?" have a different logical status than, say, the question, "What is schizophrenia?" That a diagnosis of schizophrenia, for instance, implies a hypothesization of an inner structure or state which is but inadequately described by an operational definition, and that such taxonomic conceptualization is a scientifically valuable activity has been convincingly argued by Meehl (1959). Although our present day conception of schizophrenia is riddled through with indeterminacies, we would be hard put to do without it—clinically or scientifically. The concept of normality suffers from the same type of inadequacies, only more so, since it belongs to a far less imposing nomological network than does schizophrenia.

In the case of both schizophrenia and normality we are asking legitimate questions to which, unfortunately, no sufficiently complete or specific answer can be given *at this time*. But scientific theory as it is understood today (Hempel, 1952) accepts vagueness or open endedness in the real definitions (i.e., definitions involving a statement of the essential nature or characteristics) of concepts. We should understand the question, "What is normality?" not as a demand for an unequivocal definition (as Freides implicitly does) but as a request for an empirically based specification of the *indicators* (prevalence of control factors? level of corticoseptal integration? degree of reasonableness? appropriateness of autonomic arousal? etc.) which may be found

to reflect with varying probability the existence of this hypothetical state. During the early stages of a science, the specification of the meaning of concepts is, as Kaplan (1946) notes, "a provisional one, both as to the indicators included and the weights associated with them."

The definition of normality, the hunt for the indicators of its essential characteristics, is thus construed as a processive affair marked by high responsivity to new data. An illustration incorporating some suggestive data might be of interest. To begin with, let us consider a first temporary indicator of normality such as nonexposure to psychiatric diagnosis and treatment. As an operational definition of normality, this criterion would obviously be grossly inadequate as would many others. Our gambit rests on the hope that it may eventually enable us to lift ourselves by our bootstraps (Cronbach & Meehl, 1955) by calling attention to other "purer" criteria which correlate (imperfectly) with it and which may in turn possess greater validity than our original indicator does. Many of our more successful constructs in psychology (e.g., Binet IQ) have evolved in this way. Using such a first indicator we would orient ourselves to discovering what other indicators (and characteristics by implication) compose a matrix of significant relatedness. It is clear at this point that this approach, in contrast to the idealist or average views of normality, would lead us to an empirically based conceptualization of normality having the essential character of a theoretical construct. We might pause to ask, how much reliable data do we happen to have concerning these "nonexposed" (or otherwise discriminated) normals? Surprisingly little. From reading the psychological literature a Martian might be led to believe that the proportion of schizophrenics in the United States is 80% instead of .8% or that the nonexposed represent only some 10% of the population when they should number closer to 90%.

A perusal of the literature dealing with such loosely defined normals leaves one major impression: there is an unexpectedly high distribution of "pathogenic" traits, histories, behaviors, in this population while by contrast many of the allegedly normal (ideal-adjusted) characteristics we have been taught to expect

are by no means typical of it (Lapousse & Monk, 1958; Renaud & Estess, 1955; Schofield & Balian, 1959). For our purposes this would suggest that quite a few of the admirable, clearly nonpathogenic traits and behaviors with which we have traditionally invested normality relate poorly to this concept as defined (preliminarily) by the nonexposure and other similar indicators. Parenthetically, this represents an attack on the same idealist-adjusted definition of normality that Freides finds unsatisfactory—*without, however, eliminating normality as an object of scientific search*. On the contrary, research on normality should be quickened and stimulated by such findings, presuming they stand up under more systematic investigation. Are there, for instance, suppressor-control variables which override the effect of such pathogenic events (Schofield & Balian, 1959), and are *they* the indicators of normality? Or will the answer involve a favorable interaction between certain constitutional factors and patterns of personal history? We certainly do not know the answer at this time. There can be little doubt, however, that the concept of normality constitutes an object of search that is not only scientifically appropriate but also of great moment to psychology. Such research emphasis may well be a factor in helping to precipitate a much needed shift away from the basic orientation to pathology that characterizes many of our efforts today. Clinical psychology needs to reacquaint itself with its most unique and natural subject matter—normal man. It is unfortunate, as Sanford (1958) has pointed out, that "we . . . talk much more freely about symptoms of illness than about symptoms of health or symptoms of resiliency or of strength or of spontaneity or of creativity" (pp. 82-83).

Little has been said in this rejoinder concerning the aspect of the problem of normality which bears Freides' strongest criticisms, the evaluative component of normality. Our approach simply sidesteps this argument altogether. Normality as here conceived involves absolute values or cultural biases *no more and no less* than does, e.g., our concept of schizophrenia. Certainly normality is culture bound, but only to the extent that all concepts are "bound" to the phenomena they describe or

explain. Likewise, what kind of value judgment would be involved in the empirical determination that, e.g., a specific suppressor-control variable constitutes an essential characteristic of normality? As far as we can see the only values involved in this enterprise are the same that motivate all scientific research and as such they are rightly accepted as neutral with reference to the content of science.

Although the bulk of this paper has been devoted to a defense of a conceptualization of normality, we might briefly consider the alternative that Freides champions, i.e., that we concern ourselves instead with the specific abilities and limitations that individuals demonstrate under specific circumstances. This point of view has a long history in psychology as the specificity theory of personality, and as such has been much discussed over the years. Some general comments, however, seem appropriate here. First, in addition to seeking the highest possible accuracy in the explanation and prediction of behavior, science also attempts to be maximally comprehensive and parsimonious. Freides' approach represents an admission of failure with respect to these desiderata. Secondly, the two approaches are by no means incompatible. Classifying a person as "normal" does not by any stretch of the imagination preclude an appreciation of his limitations under certain circumstances. It can easily be conceived that broadly specified situational variables may play a significant role in behavior prediction equations. To what extent this is true we do not know. It is further possible, once this typological approach

has begun to yield some fruits, that we will be able to shift to systematic studies of the processes which favor normal psychological growth. It is clear, however, that psychology would be poorly served if we gave up our search for the meaning of normality—a potentially powerful explanatory and predictive construct.

REFERENCES

- BECK, L. W. Construction and inferred entities. In H. Feigl & May Brodbeck (Eds.), *Readings in the philosophy of science*. New York: Appleton, 1953.
- CRONBACH, L. J., & MEEHL, P. E. Construct validity in psychological tests. *Psychol. Bull.*, 1955, 52, 281-302.
- FREIDES, D. Toward the elimination of the concept of normality. *J. consult. Psychol.*, 1960, 24, 128-133.
- HEMPER, C. G. Fundamentals of concept formation in empirical science. *International encyclopedia of unified science*, Vol. II, No. 7. Chicago: Univer. Chicago Press, 1952.
- KAPLAN, A. Definition and specification of meaning. *J. Phil.*, 1946, 43, 281-288.
- LAPOUSSE, REMA, & MONK, MARY A. An epidemiologic study of the behavior characteristics in children. *Amer. J. publ. Hlth.*, 1958, 48, 1134-1144.
- MEEHL, P. E. Some ruminations on the validation of clinical procedures. *Canad. J. Psychol.*, 1959, 13, 102-128.
- RENAUD, H., & ESTESS, F. Life history interviews with 100 normal American males: "Pathogenicity" of childhood. *Amer. Psychologist*, 1955, 10, 371.
- SANFORD, F. H. Psychology and the mental health movement. *Amer. Psychologist*, 1958, 13, 80-85.
- SCHOFIELD, W., & BALLAN, LUCY. A comparative study of the personal histories of schizophrenic and nonpsychiatric patients. *J. abnorm. soc. Psychol.*, 1959, 59, 216-225.

(Received June 6, 1960)

THE EFFECTS OF TWO VERBAL TECHNIQUES ON THE EXPRESSION OF FEELINGS¹

GUSTAV LEVINE²

Teachers College, Columbia University

A problem frequently encountered in psychotherapy is the client's use of impersonal, nonemotional statements. The therapist's requests for statements referring to the client's feelings are frequently met with descriptions of situations or impersonal observations, even though the client is highly motivated to cooperate. Therapists therefore frequently attempt to behave in a manner that facilitates their client's emotional expression.

Rogers, in his early writings, stated his observation that reflection of feelings results in the immediate expression of further feelings by the client (Rogers, 1942, p. 158). Studies of recorded therapy sessions which utilized categories of clients' expressions of feelings and therapists' reflections of feelings, have not examined the specific sequence of reflection and expression of feelings (Bergman, 1951; Seeman, 1949; Snyder, 1945). Verplanck (1955), in an experiment involving a social rather than a therapeutic situation, found that paraphrasing, a technique which is descriptively similar to reflection, increased the class of statements paraphrased. Although feelings were not paraphrased, but rather statements of opinion, the results are encouraging to a hypothesis that reflection of feelings increases expression of feelings. In the Verplanck study, paraphrasing of the class of responses to be increased was the only response given by the interviewers. This restricted attention to one class of response may

have been the factor which reinforced this response (rather than the specific paraphrasing technique), implying that any technique involving restricted attention to one class of response would reinforce that class. On the other hand, the specific technique (paraphrasing) may be reinforcing in the same way that approval can be reinforcing (Murray, 1956). The contention that restricted attention is the significant factor is supported by the work of Salzinger and Pisoni (1958, 1960), who found that simple statements such as, "I see," "Yeah," "Uhha," "Mmm-hmm," etc., could act as positive reinforcers of expression of feelings.

The present study involves a comparison of the specific technique of reflection of feelings with a simple undifferentiated vocalization, "Mm-hm," both applied as the only response given, in separate interviews, and occurring only after expressions of feeling. In such a pair of interviews both interviewer techniques would constitute restricted attention to a class of response, and any special advantage of the more complex response could express itself in more effective reinforcement.

It was hypothesized that reflection of feelings results in greater expression of feelings than does an undifferentiated vocalization ("Mm-hm").

The subjects were 30 male undergraduate students, living at college, who volunteered and were paid for their participation.

The experiment was explained as a study of feelings about school life. The experimenter stated that he was "interested in what you are experiencing" as a student. The subjects each received two similarly structured interviews, which differed only in the way in which the experimenter responded to expression of feel-

¹ This paper is taken from portions of a thesis submitted to the Department of Clinical Psychology, Teachers College, Columbia University, in partial fulfillment of the requirements for the PhD degree. The author wishes to express his appreciation to his Chairman, Joel Davitz, for his aid and encouragement.

² Now at the Creedmoor Institute for Psychobiologic Studies.

ings by the subjects (reflection in one, "Mm-hm" in the other).

The interviews were scheduled one week apart and tape recorded. One half of the subjects received one condition first, the other half receiving the other condition first.

The interviews were rerecorded with the interviewer's responses omitted, and with the tape divided into one-minute segments of talk by the subjects. The presence or absence of expression of feelings in a single minute was determined by six psychologists. These judges used a detailed set of instructions given to them by the experimenter as a frame of reference for their judgments.³

The basic data for the testing of the hypothesis was the number of minutes containing expressions of feeling.

The one-minute segments were rated for presence or absence of feeling by both the experimenter and a judge. The agreement was high, a phi coefficient of .92 having been obtained through a transformation of the computed chi square value of 428.9.

The average number of minutes of feeling in each of the two interview conditions was computed. In the reflection interviews there was an average of 10.07 minutes containing expression of feelings, and in the "Mm-hm" interviews there was an average of 9.83 minutes containing expression of feelings. A *t* test of the difference yielded a nonsignificant *t* of .23.

The results indicate that under the conditions of the present experiment there is no difference in effectiveness between reflection

³ The instructions are in the appendix to the thesis. The thesis can be obtained on microfilm from University Microfilms: 313 North First Street; Ann Arbor, Michigan. Order L.C. Card No. Mic. 58-2237, remitting \$2.00.

of feelings and the undifferentiated vocalization "Mm-hm," as techniques for increasing expression of feelings. The additional factor of clarification of feelings through re-expression in different words does not seem to increase the frequency of expression of feelings beyond that obtained with any technique which responds only to feelings.

The experiments of Verplanck (1955) and Greenspoon (1955), indicate that the two techniques can each be effective compared to no technique (operant level), but there was no control in the present experiment with which to make such a comparison.

REFERENCES

- BERGMAN, D. Counseling method and client responses. *J. consult. Psychol.*, 1951, 15, 216-224.
- GREENSPOON, J. The reinforcing effect of two spoken sounds on the frequency of two responses. *Amer. J. Psychol.*, 1955, 68, 409-416.
- MURRAY, E. J. A content-analysis method for studying psychotherapy. *Psychol. Monogr.*, 1956, 70(13, Whole No. 420).
- ROGERS, C. R. *Counseling and psychotherapy*. Boston: Houghton Mifflin, 1942.
- SALZINGER, K., & PISONI, STEPHANIE. Reinforcement of affect responses of schizophrenics during the clinical interview. *J. abnorm. soc. Psychol.*, 1958, 57, 84-90.
- SALZINGER, K., & PISONI, STEPHANIE. Reinforcement of verbal affect responses of normal subjects during the interview. *J. abnorm. soc. Psychol.*, 1960, 60, 127-130.
- SEEMAN, J. A study of the process of nondirective psychotherapy. *J. consult. Psychol.*, 1949, 13, 157-168.
- SNYDER, W. U. An investigation of the nature of nondirective psychotherapy. *J. gen. Psychol.*, 1945, 33, 192-223.
- VERPLANCK, W. S. The control of the content of conversation: Reinforcement of statements of opinion. *J. abnorm. soc. Psychol.*, 1955, 51, 668-676.

(Received April 29, 1960)

BRIEF REPORTS

ANXIETY IN VERBAL BEHAVIOR: AN INTERCORRELATIONAL STUDY¹

MERTON S. KRAUSE

University of Michigan

Several objective measures of anxiety in verbal behavior have been proposed in the last 3 decades. Their claims to validity are relatively weak, generally resting upon their apparent reasonableness. If they appear to reflect the same "inner state" and so yield highly correlated measurements on the same behavior samples, their claims would be stronger. They might be said to show concurrent validity.

Ten-minute recorded samples from the therapy sessions of 15 hospitalized male mental patients were studied. Eight purported measures of anxiety were applied to each patient's verbal response in the 15 protocols. The measures were (a) number of words spoken, (b) number of words/number of inspirations, (c) number of verbs/number of adjectives, (d) latency of the response, (e) number of references to the interviewer, (f) and (g) number of speech disruptions as measured by Mahl (1956) "non-ah" ratio and Dibner (1956) Cue Count I and (h) rate of speech.

An intercorrelation matrix was computed for each of the 15 protocols to study the amount of individual differences, and then the matrix of median intercorrelations was derived. The average level of correlation in this later matrix was

.06, and so no general convergence of the several measures is evident. If these measures do possess some degree of concurrent validity it was not uniform enough over persons to appear in our sample. The cluster pattern in the matrix of median intercorrelations does suggest some convergences in our set of measures. As might be expected by inspection of the measures themselves, the two speech disruption measures were highly correlated ($r_{67} = .91$), while verbs/adjectives and number of words were moderately loaded on the same factor (about .42 and .24, respectively). This pattern of Measures *f*, *g*, *c*, and *a* held up for about half of the protocols. In at least two protocols, however, this pattern clearly disintegrated.

The average results have a very large sampling error over persons. The average interquartile range for the median values was .45. Thus, there are marked individual differences in the level of intercorrelation among the various verbal anxiety measures. This implies that different measures may be valid for different persons and that what measurement values indicate anxiety or nonanxiety may also be idiosyncratic. These results suggest that verbal measures are not going to be any less troublesome to validate than are physiological measures of anxiety.

REFERENCES

- DIBNER, A. S. Cue counting: A measure of anxiety in interviews. *J. consult. Psychol.*, 1956, 20, 475-478.
MAHL, G. F. Disturbances and silences in the patients' speech in psychotherapy. *J. abnorm. soc. Psychol.*, 1956, 53, 1-15.

(Received July 1, 1960)

¹ This study was supported under Grant M-516 C-7 from the National Institute of Mental Health, E. S. Bordin, Principal Investigator.

An extended report of this study may be obtained without charge from Merton S. Krause (2343 Auburn Avenue; Cincinnati 19, Ohio) or for a fee from the American Documentation Institute. Order Document No. 6642, from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance \$1.25 for microfilm or \$1.25 for photocopies. Make checks payable to: Chief, Photoduplication Service, Library of Congress.

VERBAL AND PERCEPTUAL COMPONENTS IN WISC PERFORMANCE AND THEIR RELATION TO SOCIAL CLASS¹

JOHN B. MARKS

Veterans Administration Hospital, American Lake, Washington

AND JAMES E. KLAHN

Tacoma Public Schools

Most investigators have found measured intelligence positively related to social class and have found this relation closer with verbal rather than nonverbal materials. With the WISC, however, Estes (1953) found higher status superiority only at a 7-year-old level and not at the 10-year-old level. Moreover, she found no consistent pattern differences of subtests between her upper social group and her lower group.

The present study relates social class to two WISC measures of verbal-nonverbal difference: (a) Verbal IQ-Performance IQ, and (b) the difference between weighted scores of subtests highly loaded on Cohen's (1959) verbal factors and weighted scores of subtests highly loaded on the perceptual factor. This latter was the mean of Information, Comprehension, Similarities, and Vocabulary minus the mean of Picture Completion, Block Design, and Object Assembly.

Subjects were 211 primary school children divided by age and sex into four groups. Both younger groups were within 6 months of their eighth birthday at testing while the older groups were within 6 months of their eleventh birthday. All had been tested because of some school difficulty but children with IQs below 70 were eliminated in sample selection. Mean IQs of the sample were close to population means.

From information about the father's occupation each subject was assigned to an occupational class group ranging from 1, casual laborer, to 9, business leader. The r between independent rat-

ings of the two authors was .94 and means of the two ratings were used.

Occupational ratings correlated positively with IQs in both the younger and the older groups. Though the correlation in the younger group is higher it is not significantly so. On the other hand, the girls show a substantially higher correlation than do the boys. For verbal and full-scale IQs this difference is significant, .42 compared to .19 for verbal and .45 to .17 in the full-scale.

The two measures of verbal-nonverbal difference were tested for their relation to occupational level, both directly through correlations, and by contrasting the difference measures for unskilled and semiskilled labor children with those for white collar children. The difference between verbal and performance IQs was in the expected direction but not significant. The difference in the factor-derived measures had a significant r of .16 ($p < .05$, $N = 211$) with occupational level and the white collar children showed a greater verbal superiority ($t = 2.41$, $p < .01$, 113 df).

These results are consonant now with results using other instruments. On the WISC both younger and older groups show a correlation of IQ with occupational class, and this correlation is higher when verbal materials are used than it is for nonverbal. The closer relation between occupational level and IQ among girls than among boys may stem from the higher peer value which girls put upon middle class verballity.

REFERENCES

- COHEN, J. The factorial structure of the WISC at ages 7-6, 10-6, and 13-6. *J. consult. Psychol.*, 1959, 23, 285-299.
ESTES, BETSY W. The influence of socioeconomic status on the WISC: An exploratory study. *J. consult. Psychol.*, 1953, 17, 58-62.

(Received July 26, 1960)

¹ An extended report of this study may be obtained without charge from John B. Marks (Mental Health Research Institute; Fort Steilacoom, Washington) or for a fee from the American Documentation Institute. Order Document No. 6643, from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., receiving in advance \$1.25 for microfilm or \$1.25 for photocopies. Make checks payable to: Chief, Photoduplication Service, Library of Congress.

SENSORY DEPRIVATION AND ITS RELATION TO PROJECTION¹

MALCOLM H. ROBERTSON AND ROBERT C. MARTIN

University of Florida

The study was designed to test the hypothesis that sensory deprivation lowers the threshold for projection. To test this hypothesis, a control group and an experimental group of 10 subjects each were used, half male, half female. The control subjects received no deprivation and were tested individually for projection using a modified autokinetic technique. The technique consisted of presenting the subject with a dim source of light approximately 1 mm. in diameter at a distance of 9 feet. The subject was told to watch the moving pinpoint of light and, when it went off, to report what it suggested, or looked like, or made him think of. There were 12 1-minute trials with a 2-minute rest period after the sixth trial.

The experimental subjects were exposed to sensory deprivation for 3 hours and then tested immediately with the autokinetic technique in the same manner as the control subjects. In the deprivation condition, each subject wearing opaque goggles, cotton mittens, and cardboard cuffs, was placed on a bed with his head inside a foam rubber lined box.

The two groups were compared in terms of the number of responses as well as the number of stimulus-bound responses, original responses, and popular responses. Stimulus-bound responses were those that referred solely to the movement or direction of movement of the light. An original response was one that was given only once, by only one person, and in addition was considered by the two investigators to be very novel or unusual. A popular response was one that was found in several records, usually occurring more than once in a record.

It was expected that the deprivation group would show more projection (greater productivity, larger number of original responses, and fewer popular and stimulus-bound responses) than the control group. Differences between the two groups were not statistically significant.

¹ An extended report of this study may be obtained without charge from Malcolm H. Robertson (Department of Psychology, University of Florida; Gainesville, Florida) or for a fee from the American Documentation Institute. Order Document No. 6644 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance \$1.25 for microfilm or \$1.25 for photocopies. Make checks payable to: Chief, Photoduplication Service, Library of Congress.

(Received August 8, 1960)

PSYCHIATRIC OUTPATIENT PERSONALITY PATTERNS¹

MARY HELEN TATOM
Spring Grove State Hospital, Baltimore

In a previous study (Tatom, 1958) five patients were selected on the basis of medical diagnoses as representatives of each of four commonly occurring nosological entities: obsessive-compulsive, hysteric, anxiety state, and outpatient schizophrenic. They were rated by their respective individual therapists on each of 67 personality variables, and the resulting scores intercorrelated. The 20 × 20 matrix of person-to-person correlations was factored by Thurstone's centroid method to yield four primary and two second-order factors, tentatively identified with clinical syndromes.

To test the stability of these patterns, 2 reference individuals for each of the four primary factors were inserted into a matrix with 16 new patients, selected to represent equally the original four clinical diagnostic entities. A factor analysis paralleling the first was carried out. The position of reference individuals with respect to primary factors in the two analyses is given in Table 1. Trait patterns of structurally corresponding fac-

tors in the two analyses suggested that syndromes were replicated in the second analysis, though overlap of specific items was not extensive (e.g., two internally consistent clusters of clinically schizoid traits had in common only *seclusiveness* and *poor social adjustment*). Primary factors were tentatively identified with the respective factors of the original analysis as follows: *outpatient paranoid schizophrenic*, *conversion hysteric*, *socially mature personality*, and *passive-dependent personality vs. antisocial personality*. The latter was less clearly defined by reference individuals than were the other three. Patients in both analyses were grouped largely on the basis of second order factors, rather than highly correlated primary factors. Factor and trait patterns tended to confirm two second-order factors: *schizothymia vs. cyclothymia*, and *uncontrolled emotionality vs. overcontrolled emotionality*.

Psychiatric and factorial classification of patients did not agree in either analysis; syndromes isolated factorially corresponded only roughly to the nosological entities represented by patients in both groups of patients analyzed.

REFERENCE

TATOM, MARY H. A factorial isolation of psychiatric outpatient syndromes. *J. consult. Psychol.*, 1958, 22, 73-81.

(Received August 25, 1960)

TABLE 1
PROFILES OF FACTORS IN REFERENCE INDIVIDUALS

Reference Individual	Factor							
	A	A ₂	B	B ₂	C	C ₂	D	D ₂
O-5 (A)	.60	.66	-.20	.00	.02	-.08	-.26	.13
O-3 (A)	.67	.87	-.03	.06	.00	-.05	-.05	-.07
A-1 (B)	-.20	-.27	.49	.71	.11	.28	-.01	.18
A-4 (B)	.00	-.02	.57	.64	.01	-.01	.04	-.14
H-3 (C)	.02	-.02	.01	.06	.69	.71	-.03	.00
A-2 (C)	-.08	-.39	.44	.01	.58	.40	.26	.51
H-2 (D)	-.26	-.64	-.11	.38	-.02	.39	-.49	.37
H-4 (D)	.13	-.39	.00	.52	-.04	.42	-.49	.23

IDENTIFICATION IN TERMS OF PERSONAL CONSTRUCTS: RECONCILING A PARADOX IN THEORY¹

ROBERT E. JONES

Veterans Administration Hospital, Danville, Illinois

In this study, identification is defined as perceived similarity of self and others, experienced in terms of personal constructs. The author accepts the definition of phenomenological psychologists and the thinking of psychoanalysts, such as R. P. Knight, who emphasize identification as a relationship rather than a process, and as a perceived rather than as an actual similarity.

The psychology of personal constructs, the theoretical system developed by G. A. Kelly, is a perceptual approach to the prediction and explanation of human behavior. One's personal constructs are the vehicles, verbally expressed, by which one anticipates the behavior of others and guides his own behavior. Constructs are dimensions defined by terms which the perceiver accepts as opposites or contrasts. The way in which two persons are seen as alike yet different from a third would be a construct. A form of the Role Construct Repertory Test is employed to measure identification with "significant others."

Identification with male figures in the Repertory test was taken as the central measure in the present research. Repeat-test reliability was established in the high .80s. The subjects were 36 hospitalized males with mild or moderate psychiatric disorders, and a control group of 36 normal males, matched on age and education.

The central hypotheses were two: (a) neuropsychiatric (NP) patients more often than normal adult males will either overidentify or under-

identify with personally significant male figures; and (b) the personal construct matrices of NP patients will be simpler than those of controls.

The hypothesis of underidentification is supported only at the 10% level, by one-tailed *t* test. The hypothesis of overidentification is significant at the 1% level: NPs are more likely to see others as extremely like the self than are normals. Also as predicted, the idiographic factors required to "explain" the interpersonal matrix are significantly fewer (at the .05 level) for NPs than for the controls. For NPs, but not for normals, the more simple the factor matrix the more fully it is explained by a value construct (.01 chi square significance).

The major contribution of the Personal Construct approach to identification theory lies in the reconciliation of previous theories. Both E. H. Erikson and O. H. Mowrer are partly right. Both over- and underidentification are common badges of maladjustment. Both are associated with a common defect—a factorially simple, value-laden system of constructs. Dimensions of perception permitting useful discrimination become inoperant. With the construct system compelling polarization into "good guys and bad," we see, with Sanford, how "desperation" promotes classical identification of the all-or-none variety. Identification with a hated object tends to be unconscious and accomplished by introjection. Identification with an idealized object tends to be conscious and achieved by assimilative projection. Sanford's "identification proper" is the unconscious type, always desperation motivated. Either type of identification, whether excessive or deficient, can be explained in terms of an oversimplified construct system, preempted by the value dimension.

The findings support Rinder and Campbell's contention that both over- and underidentification reflect the same neurotic dynamic: in their terms, "undue reaction-sensitivity"; in ours, undue perceptual restriction to the value dimension.

¹ An extended report of this study may be obtained without charge from Robert E. Jones (Psychology Service, Veterans Administration Hospital, Danville, Illinois) or for a fee from the American Documentation Institute. Order Document No. 6646 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance \$1.75 for microfilm or \$2.50 for photocopies. Make checks payable to: Chief, Photoduplication Service, Library of Congress.

This paper draws on data collected when the author was engaged in doctoral research at Ohio State University.

(Received September 22, 1960)

A STUDY OF READING DISABILITY

RICHARD H. WALTERS,¹ MALLE VAN LOAN, AND IRENE CROFTS²

University of Toronto

METHOD

Subjects

Ss were 58 Grade 3-6 boys from a single suburban school. They were selected from a larger group of 86 boys who were of average intelligence (IQs between 88 and 112) and who were, according to their medical histories, free of eye anomalies, hearing losses, and behavior problems. Of the 86 boys, 20 had reading ages that were at least 1 year below their mental ages; these boys were regarded as retarded readers. Eighteen boys whose reading ages were at least 1 year above their mental ages were regarded as advanced readers. From the remaining 48 boys, the 20 boys with the smallest discrepancies (less than 6 months) between reading age and mental age were selected as average readers.

Reading ages and mental ages were available in the school records; since the reading tests and intelligence tests had, in some cases, been administered several months apart, the mental ages were corrected to correspond to the time at which the reading test was given. Unfortunately, the tests administered varied to some extent from grade to grade. However, since the administration of further tests would have too greatly disrupted the school timetable, the available indices were accepted as reasonably adequate bases for group selection. The tests given were, in fact, very similar, e.g., the Otis group test and the Dominion Test of Learning Capacity as measures of intelligence and the Gates Vocabulary, Shenell Vocabulary, and Dominion Silent Reading Tests as measures of reading ability.

Apparatus and Tests

Two perceptual tests⁴ developed for the Cerebral Palsy Project of the Department of Psychology, University of Toronto, were included in the test battery. The Steer-Allen Figure-Ground Confusion Test consists of 15 cards, on each of which is depicted an object whose outline is interrupted by the background. In order to identify the object, S must free the figure from the confusing background detail. Since this test was developed for studies of children younger than those used in this study, it was employed primarily as a rapport building test at the commencement of

promising results for boys, but negative results for girls. Mimeographed copies of this study may be obtained from the senior author.

⁴ The authors are indebted to H. O. Steer both for permission to use these tests and for his helpful advice on this study.

The psychoanalytic theory of reading disability (Blanchard, 1946; Fenichel, 1937; Strachey, 1930), which is closely related to Freud's (1924) theory concerning hysterical blindness, has received relatively little attention from research psychologists in spite of its popularity among psychoanalytically oriented therapists. In a recent outline of this theory, Jarvis (1958) has drawn attention to three factors that supposedly characterize the retarded reader: an avoidance of looking, the problem of aggression, and faulty identification mechanisms. Jarvis regards the "active part of looking" as creating the major difficulty for the retarded reader; the counterpart of this activity outside the school "is an inability to identify predominantly with one's own sex" (p. 468).

The studies reported in this paper were suggested by the psychoanalytic theory of reading disability. Since it is the alleged sexual significance of reading that, according to psychoanalysts, gives rise to fear or avoidance of looking, it was hypothesized that retarded readers would show greater hesitation in looking at a sexual object than would either average or advanced readers, and that, through generalization, this "avoidance" would also be evident in perceptual tasks involving identification of nonsexual objects. It was further hypothesized that retarded readers would display hostility toward the same-sex parent, a condition that might reflect both aggression and "faulty identification mechanisms."³

¹ The authors wish to express their appreciation to the Scarborough Board of Education and to the staffs of the R. H. King Collegiate and Corvette Schools, especially to Mathilde Ziehr and James Phillips, for their co-operation in the studies reported in this paper. They are also grateful to Marian B. Hyatt for her assistance in testing.

² Now at the Child Guidance Clinic of Greater Winnipeg.

³ A preliminary study of the relationship between reading achievement and identification produced



FIG. 1. Multiple-choice apparatus, showing nude doll used as test object.

testing. The Steer-Beatty Closure-Threshold Test, a more suitable test for children in Grades 3-6, consists of 12 sets of five cards, each card being composed of dots. Each set forms a graduated series of representations of the same object. The degree of clarity of the representations is a function of the relative spacing of dots that compose the figure and dots that compose the background. The cards in each set are presented consecutively in an order that makes the identification of the object (figure) increasingly less difficult. These perceptual tests were included because they appeared to require "active looking" in the sense in which this term is used by psychoanalysts.

A multiple-choice apparatus (Figure 1) was used for the major test of the "fear of looking" hypothesis. A box, approximately 3 feet long and 1½ feet high, was separated into four compartments. Each compartment was fronted by a separate door which opened easily. The back of the box also opened to allow the experimenter (*E*) to insert an object into any one of the compartments. The doors and the back of the box were connected with a buzzer and an electric timer. When the box was completely closed, back and front, the electrical circuit was completed and both the buzzer and timer were set in operation. Opening any one of the doors interrupted the circuit and shut off both the buzzer and the timing mechanism. This apparatus provided an automatic recording of the interval between *E*'s signal for *S* to respond (buzzer) and *S*'s opening of one of the doors to look at a hidden object. Three objects were used during testing: a female Dutch doll of a conventional type (neutral object), a nude male doll with a penis (the type sometimes employed by psychoanalytically oriented child psychiatrists), and clothed boy doll (a second neutral object).⁵

Since psychoanalysts have stressed the importance of "underlying" unconscious psychological determinants, and some researchers (e.g., Friedman, 1952)

⁵ The choice of a nude male doll as the test object was based on psychoanalytic symbolism (Fenichel, 1937; Freud, 1953; Strachey, 1930) concerning the act of looking and the mastery of reading. The play therapy dolls were lent by Alice Moulby of the York Township Child and Adolescent Guidance Clinic.

have claimed that these can be diagnosed only through the use of projective techniques, a brief picture-story test was added to the test battery. Four pictures were included:

1. A boy is shown turning his back on an older male figure, who is walking away in the opposite direction.

2. A boy is shown looking into a bathroom. An arm of a taller figure protrudes from the shower curtain, and water can be seen coming from the rose of the shower.

3. A boy stands in front of an older seated male figure, who is reading a newspaper. The boy is gesticulating with arms outstretched.

4. A boy is shown looking into a room in which an adult male and female are embracing.

Pictures 1 and 3 were chosen to test hostility toward the father; Pictures 2 and 4 to test fear and avoidance of looking.

An attempt was made to assess identification by a modified version of Fiedler's (1958) Assumed Similarity to Others (ASo) Test. This test, however, proved to be beyond the comprehension of the younger children in the study and, consequently, the results were of little value. A simple test of parent preference, described below, was also given.

Procedure

S was brought to the experimental room by a female *E* and was seated at a desk facing the discrimination apparatus. *E* seated herself to the rear of the apparatus about 4 feet away from *S*.

E first presented the Figure-Ground Confusion Test, using the following instructions: "I am going to show you some pictures in each of which a thing is hidden. I want you to tell me what that hidden thing is." *S* was given a trial run with a card representing a bird, after which *E* pointed out the detail of the bird. The remaining pictures were then presented one at a time. No time limit was set. If *S* said that he could not find the object, *E* simply presented the next card. *S*'s responses were recorded on data sheets.

S was now asked to stand in front of the discrimination apparatus and was given the following instructions:

Here I have this doll [*E* held up the Dutch doll] and I am going to hide it in one of these boxes. I want you to see if you can find it. You can do this by opening one of these doors [demonstrated] as soon as you hear the buzzer. When you hear the buzzer you can open the door in which you think I have hidden the doll. Don't close the door again until I tell you to.

S was given three trial runs with the Dutch doll, then the experiment proper was begun. The doll was placed into the various compartments from the back of the apparatus in a predetermined random sequence. *S* was given 10 trials with this doll. The

⁶ These pictures form part of a series developed by Albert Bandura of Stanford University.

latency and choice of box in each trial were recorded by *E*.

E now took out the nude father doll, holding it up in the air in full view of *S*. *E* said: "Look at this doll! Now I am going to do the same thing with this doll, and I want you to do exactly as you did before." Again the doll was placed, in a predetermined random sequence, into each of the compartments. Ten trials were given with the father doll, and the results were recorded as before.

Finally, *E* took the clothed boy doll. Holding it up, *E* said: "Now I am going to use this doll, and I want you to do exactly the same as before." This time the doll was placed in each compartment in an orderly sequence: 1-2-3-4, 1-2-3-4. The purpose of this final set of trials was twofold. It was thought that the subsequent presentation of a neutral object might reduce any emotional upset produced by the nude doll; in addition, through the use of a regular sequence, it seemed likely that *S* would finish up by making some "correct" responses, so reducing possible feelings of failure. The plan was to continue hiding the boy doll until *S* had made two successive correct responses. Within the time limit imposed by the school schedule, this was not possible in all cases.

The design of the experiment would have been improved if presentations of a neutral doll and the nude doll had been made in random order. However, to minimize the possibility of emotional upset and of subsequent parental complaints about the use of a nude figure, it was thought wiser to buttress the presentations of the nude figure with preceding and subsequent presentations of neutral figures.

S was now taken to a second female *E*, who had not been associated with the presentation of the nude figure. *E* seated *S* beside her and took out a quarter. She said:

Now I want you to imagine that you are going on a long, long trip. You can go with your mother or your father, but you cannot go with both. I am going to toss this coin. "Heads" you go with your father, "tails" you go with your mother. You call.

A preliminary trial was given to insure that *S* complied with instructions. The instructions were then carefully repeated, and a further trial was given. *S*'s response was recorded on each trial.

E now gave the picture-story test. *S* was instructed:

I want you to make me up a story about this picture. Tell me what the little boy is doing, what led up to this, how the little boy is thinking and feeling, and how it will all turn out.

The only probes given were repetitions of parts of the original request or variations on these. *S*'s stories were recorded on tape and later transcribed.

The abortive ASO test was then administered. Finally, *E* asked the *S* to stand at a point in the room approximately 8 feet away from her and presented the Closure-Threshold Test. The sets of cards were presented in a standard order. *S* was shown the first (most difficult) card in a set and was asked: "What

TABLE 1

DISTRIBUTION OF ERROR-FREE RECORDS ON THE FIGURE-GROUND CONFUSION TEST

Errors	Advanced Readers	Average Readers	Retarded Readers
Present	7	7	2
Absent	11	13	18

Note.—Chi square = 4.797; $p < .10$.

does this look like to you?" If *S* failed on this card, he was shown the next card in the series and was asked what it looked like. The procedure was continued until *S* responded correctly or until all five cards in a series had been presented. On Cards II to V, if *S* said he did not know what the card looked like or that it did not look like anything, *E* said: "Perhaps you can tell me what it looks most like. What do you think it might be?" The procedure was continued until all 12 sets of cards had been shown.

The testing was now complete, and *S* was taken back to his classroom.

RESULTS

Avoidance of Looking

As expected, a large number of *Ss* obtained a perfectly correct record on the Figure-Ground Confusion Test. Consequently, a chi square analysis was performed with the data divided into two categories—errors present and errors absent. The distribution of cases among the three groups of *Ss* is given in Table 1. The results were in the predicted direction; however, the distribution could have occurred by chance 10 times in 100.

Results of the Closure-Threshold Test were assessed as follows. For each set of cards, *S*'s score consisted of the number of cards required to elicit a correct response; if *S* failed

TABLE 2

ANALYSIS OF VARIANCE BY RANKS OF RESPONSES TO THE CLOSURE-THRESHOLD TEST

Sums of Ranks			
Advanced Readers (<i>N</i> = 18)	Average Readers (<i>N</i> = 20)	Retarded Readers (<i>N</i> = 20)	<i>H</i>
429.5	580.0	701.5	6.219*

* $p < .05$.

TABLE 3

ADJUSTED GROUP MEAN LATENCIES (IN SECONDS) OF RESPONSES TO THE NUDE DOLL

Advanced Readers (<i>N</i> = 18)	Average Readers (<i>N</i> = 20)	Retarded Readers (<i>N</i> = 20)
1.417	1.744	2.044

TABLE 4

ANALYSIS OF COVARIANCE OF LATENCIES OF RESPONSES TO NUDE DOLL

Source	Adjusted <i>SS</i>	<i>df</i>	Adjusted <i>MS</i>	<i>F</i>
Between groups	3.137	2	1.568	10.181**
Within groups	8.331	54	0.154	

** $p < .001$.

to respond correctly to the fifth (the easiest) card, he was arbitrarily assigned a score of 6. These component scores were summed to provide a total score for each *S*. Total scores were then ranked for a Kruskal-Wallis analysis of variance by ranks (Siegel, 1956). Results were as predicted (Table 2).

The median latency of response to the first two dolls in the multiple-choice task was computed for each *S*. Medians were preferred to means as a measure of central tendency because of the possible presence of a single deviant response within a series of trials. Using the responses to the nude doll as the dependent variable, an analysis of covariance was carried out with responses to the Dutch doll as the covariant. Table 3 gives the adjusted group mean latencies to the nude doll, and Table 4 the results of the analysis of covariance. A test of homogeneity of regression had previously indicated that this procedure was justifiable ($F = 1.95$; $p > .05$). Differences were in the predicted direction and were significant at the .001 level. Subsequent *t* tests indicated that this result was largely due to the superior performance of the advanced readers, whose performance was significantly better ($p < .001$, one-tailed test) than that of either of the other two groups. Because all three groups of *Ss* tended to show some de-

crease in latency in response to the nude doll (undoubtedly a practice effect), a further analysis of covariance of responses to the nude doll was carried out, this time with responses to the boy doll as the covariant. Significant differences in the predicted direction were obtained ($F = 6.46$; $p < .005$).

Parent Preference Test

The results of the coin tossing test (second trial) are given in Table 5. The percentage of retarded readers showing preference for the same-sex parent was smaller than that of either of the other two groups ($p < .02$). When results from the preliminary trial were included in the analysis, the differences among the three groups were considerably reduced (Table 6).

Picture-Story Test

The stories were scored from typescripts by an assistant who did not know to which group individual *Ss* belonged. Instructions for scoring were as follows:

Story 1. Score + if taller figure is identified as the father and if the boy in the picture expresses hostility toward the father (i.e., is described as "angry," "mad," or as wishing to harm the father).

TABLE 5

DISTRIBUTION OF PREFERENCES ON COIN-TOSSING TEST: SECOND TRIAL

Preference	Advanced Readers	Average Readers	Retarded Readers
For father	13	10	4
For mother	5	10	16

Note.—Chi square = 10.561; $p < .01$.

TABLE 6

DISTRIBUTION OF PREFERENCES ON COIN-TOSSING TEST: COMBINED TRIALS

Preference	Advanced Readers	Average Readers	Retarded Readers
For father on both trials	9	7	3
For mother on one or both trials	9	13	17

Note.—Chi square = 5.339; $p < .10$.

Story 2. Score + if the boy in the picture (a) does not mention the figure behind the shower curtain, or (b) is described as unwilling to enter the bathroom because someone is already in there, or (c) is described as being afraid, ashamed, or guilty, because he has seen the figure in the shower.

Story 3. Score + if seated figure is identified as the father and if the boy expresses hostility toward the father (as for Story 1).

Story 4. Score + if the boy in the picture (a) does not describe the man and woman as kissing or making love, or (b) is described as being afraid, ashamed, or guilty because he has seen the adult figure making love.

For the purpose of analysis, Ss were regarded as showing hostility toward the father if they received a + score on either Story 1 or Story 3, and as showing fear of looking if they received a + score on either Story 2 or Story 4. Subsequent chi square tests failed to support the hypotheses being tested.

DISCUSSION

Only the "fear of looking" tests could provide crucial support for the psychoanalytic theory. One of these, the picture-story test, yielded completely negative results. The multiple-choice task, on the other hand, may be interpreted as supporting the psychoanalytic theory. However, in view of the negative results of the picture-story test, alternative interpretations must be favored.

Since the differences between advanced readers and both the other groups of Ss on the multiple-choice task were significant at the .001 level, interpretation could depend heavily upon the characteristics of over-achievers, and particularly upon information about child training practices that tend to produce high-achieving children. Unfortunately, little is known about the family backgrounds of children who are exceptionally advanced in reading. On the other hand, the studies of McClelland, Atkinson, Clark, and Lowell (1953) do give some information concerning the family backgrounds of Ss with high need achievement. It is possible that over-achieving readers fall within the larger group of individuals who are highly oriented toward achievement in general. Making this assumption, one might expect that over-achieving readers come from homes in which there is considerable stress on independence training,

in which the parents are democratic rather than autocratic, and in which conventional moral standards are not highly stressed. It is this latter factor that may be important in interpreting the results of this experiment. A child who is not deterred from peering into a compartment that, if his choice is correct, contains a nude figure with genitals may thus be the product of a home in which conventional moral standards are not stressed.

Observance of conventional moral standards in the area of sex training involves considerable emphasis on modesty (Sears, McCoby, & Levin, 1957). If the emphasis on modesty training is not so strong in the homes of high achieving children, one might expect that these children have had opportunity to see their same-sex siblings and fathers in the nude. Therefore one would not expect the over-achieving children in this study to be greatly deterred from responding quickly in a task involving a nude doll of the same sex as themselves.

McClelland et al. (1957) have found that individuals with high need achievement tend to come from homes in which the parents make strong demands for independence, including strong achievement demands. Thus, these children may not only be relatively uninhibited about sexual matters, but may also have been reinforced for exploratory behavior. This latter factor, along with low emphasis on modesty training, may be partly responsible for the obtained difference between the over-achievers and the other two groups of Ss.

Since, for all three groups of Ss, there was a decrease in median latency over the three sets of trials, it seemed possible that the inclusion of the nude doll as a test object was, in fact, of little importance, and that the results of the discrimination test merely reflected differences in learning capacity in a perceptual-motor task. This interpretation is partly borne out by Figure 2. In this figure changes in test object (Dutch doll, nude doll, boy doll) are ignored. The median response of each S on each block of 5 trials was first identified. These median latencies were then averaged to provide an indication of changes in latency over a series of 30 trials. The figure indicates that advanced readers improve

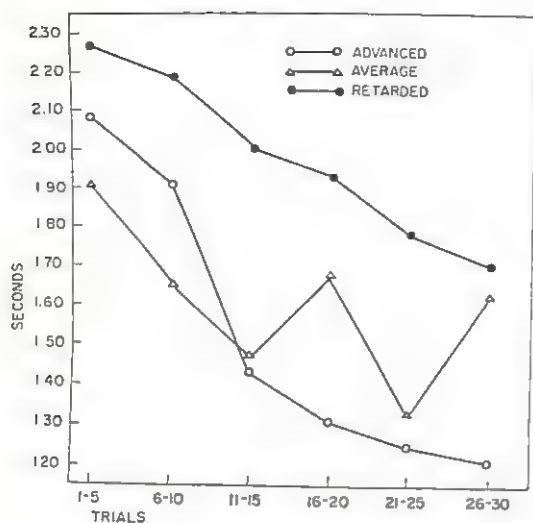


FIG. 2. Change in latencies of advanced, average, and retarded readers over a series of 30 trials, for all stimuli in the "looking" test.

in performance at a much faster rate than do retarded readers. Average readers show an initial improvement that is intermediate between that of the other two groups, then perform somewhat erratically. These results suggest the necessity for a further test in which a single test object is utilized and in which trials are continued until asymptotes are reached for all three groups.

On the two perceptual tests the retarded readers performed more poorly than both the average and the advanced readers. Once again, these results could be interpreted as supporting the psychoanalytic theory. A simpler explanation, however, is that perceptual skills are highly developed among advanced readers and poorly developed among retarded readers, and that the development of these skills is related to reinforcement of exploratory behavior by care-taking adults.

The above explanation of the findings imply that fear of looking is not a *causative* factor producing differences among the three groups of Ss in their responses to the nude doll. It is suggested that, at the most, parents who inhibit exploratory behavior are also nonpermissive in their modesty training and that, as a consequence, children who are retarded readers tend also to be sexually inhibited. In this connection it is important to remember that during a child's early years

sexual behavior largely occurs in the form of curiosity or exploratory behavior involving perceptual-motor responses.

In an attempt to integrate the findings concerning hostility and identification into a tentative theory concerning the antecedents of reading disability, the fantasy data will be ignored on the grounds that, in spite of their widespread utilization in clinical settings, such data seem to have no consistent relationship with supposedly corresponding overt responses. The coin tossing test suggests that boys who are retarded readers are relatively hostile toward their fathers.

Let us now make the further assumption, for which child training studies afford some justification, that mothers tend to be somewhat nonpermissive concerning exploratory behavior and that fathers are more variable in this respect. In this case, the amount of reinforcement which exploratory behavior receives will depend considerably upon the father's behavior patterns. Hostility to the father may then be viewed as an outgrowth of paternal nonpermissiveness and punitiveness, i.e., frustration, of exploratory behavior, of which sexual behavior is an important facet.

SUMMARY

Psychoanalysts have attributed reading disability to three, supposedly related, factors: fear and avoidance of looking; hostility, primarily toward the same-sex parent; and failure to identify with the same-sex parent.

Hypotheses suggested by psychoanalytic theory were investigated in a study in which Grade 3-6 boys were employed as Ss. Retarded readers performed more poorly on two perceptual tasks involving recognition of form than did average and advanced readers. They were slower in opening a compartment to look for a male nude doll than were the other two groups. In addition, they chose their father less often on a simple parent preference test. On the other hand, fantasy data failed to support either the "fear of looking" or the hostility hypothesis.

Some of the above results may be interpreted as supporting the psychoanalytic theory. An alternative interpretation in terms of parental conditioning of exploratory and sexual responses was nevertheless favored.

REFERENCES

- BLANCHARD, PHYLLIS. Psychoanalytic contributions to the problem of reading disability. *Psychoanal. Study Child.*, 1946, 2, 163-168.
- FENICHEL, O. The scopophilic instinct and identification. *Int. J. Psycho-Anal.*, 1937, 18, 6-34.
- FIEDLER, F. E. *Leader attitude and group effectiveness*. Urbana, Ill.: Univer. Illinois Press, 1958.
- FREUD, S. Psychogenic visual disturbance according to psychoanalytic concepts. In E. Jones (Ed.), *Collected papers of Sigmund Freud*. Vol. 2. London: Hogarth, 1924.
- FREUD, S. The interpretation of dreams. In J. Strachey (Ed.), *The standard edition of the complete psychological works of Sigmund Freud*. London: Hogarth, 1953.
- FRIEDMAN, S. M. An empirical study of the castration and oedipus complexes. *Genet. psychol. Monogr.*, 1952, 46, 61-130.
- JARVIS, V. Clinical observations on the visual problem in reading disability. *Psychoanal. Study Child.*, 1958, 13, 451-470.
- McCLELLAND, D. C., ATKINSON, J. W., CLARK, R. A., & LOWELL, E. L. *The achievement motive*. New York: Appleton-Century-Crofts, 1953.
- SEARS, R. R., MACCOBY, ELEANOR E., & LEVIN, H. *Patterns of child rearing*. Evanston, Ill.: Row, Peterson, 1957.
- SIEGEL, S. *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill, 1956.
- STRACHEY, J. Some unconscious factors in reading. *Int. J. Psycho-Anal.*, 1930, 11, 322-331.

(Received June 21, 1960)

PSYCHOLOGISTS' JUDGMENTS OF PHYSICAL HANDICAP FROM H-T-P DRAWINGS

ORVAL G. JOHNSON

Lewis County Schools, Washington

AND

FRANK WAWRZASZEK

Eastern Michigan University

The psychological characteristics of the physically handicapped person have been studied increasingly in recent years. Force (1956), Lerner and Martin (1955), Shere (1956), Whitehouse (1953), Wrightstone (1957), and Cruickshank (1955) suggest various degrees and kinds of differences psychologically between handicapped and nonhandicapped persons. On the other hand, Levy and Michelson (1952) and Wenar (1956, 1958) find no differences or only minor differences between the two groups. Berreman (1954) looks at the problem from the standpoint of the attitudes shown toward handicapped people, which he says are different from attitudes toward normals, and are likely to affect the self-images of the handicapped. Wawrzaszek, Johnson, and Sciera (1958) found no differences between handicapped and nonhandicapped children on any of 10 variables derived from the House-Tree-Person Test.

Problem

The purposes of this study are twofold:

1. To determine whether or not physically handicapped children project into their drawings any feelings about their handicaps to the extent that the handicap can be detected by psychologists through an analysis of their drawings

2. To investigate the characteristics of drawings that psychologists use to postdict which of a matched pair of children is handicapped

METHOD

The H-T-P test was administered to 37 handicapped children in the special classes of a public school. Sixteen were postpolio cases, seven cardiac, seven cerebral palsied (mild), two spina bifida, and

one each Perthes hip, muscular dystrophy, congenital deformity, slipped epiphysis, and brittle bones.

These children were selected from a slightly larger group, eliminating those who might be hampered in their drawing by impaired motor coordination. This selection was accomplished with the consultation of the physical therapist.

A control group was made up by matching each of the handicapped on the basis of age, sex, and IQ. The chronological age of each control was within 3 months and the IQ within 10 points of that of the handicapped child. Therefore, by definition, the average chronological age and IQ for both groups were the same. Actual computations showed that there was no significant difference between the two groups in CA and IQ. All but a few of the handicapped and control children had been tested with individual intelligence tests. Table 1 shows the mean, range, and SD of CA and IQ for both groups.

The H-T-P test was administered to each child in accordance with the instructions suggested by Buck (1948). The matched pairs of protocols, randomized for order of handicapped and nonhandicapped, were presented to nine psychologists with the following instructions:

These are H-T-P protocols of 37 pairs of children matched for sex, CA, and IQ. One of the children is physically handicapped and one is his nonhandicapped control. You are to judge from the protocol which is the physically handicapped child and which is the control. Please clip a slip of paper onto the protocol of the handicapped child, noting briefly the basis for your decision in each case.

TABLE 1
MEAN, RANGE, AND STANDARD DEVIATION OF HANDICAPPED AND NONHANDICAPPED GROUPS ON CHRONOLOGICAL AGE AND INTELLIGENCE QUOTIENT

Measure	Chronological Age		Intelligence Quotient	
	Handicapped	Nonhandicapped	Handicapped	Nonhandicapped
Mean	116.3	118.3	103.4	104.2
Range	6-5 to 13-6	6-6 to 13-5	63 to 128	63 to 129
SD	25.6	26.1	13.9	12.7

All of the psychologists had experience in working with children, a majority of them having done psychological work with physically handicapped children. The drawings of the Person were evaluated according to the Goodenough (1926) scale, and an MA based thereon derived for each subject. The drawings of the Person were also measured for height.

RESULTS

Of the 333 judgments (9 judges \times 37 judgments for each), 208 or 62% were correct. This proportion is significantly different (.01 level) from chance expectation. The percentages correct for the judges individually ranged from 54 to 65. None of these is significantly different from chance.

In many cases there was substantial agreement among the psychologists as to which of the protocols was that of the handicapped child. Sometimes they agreed and were right, and sometimes they agreed and were wrong. If we arbitrarily assume that more than two-thirds agreement by the psychologists is substantial agreement, then it can be said that in 27 out of 37 possible cases, or 73%, there was substantial agreement. In 19 out of the 27 cases where substantial agreement occurred, the majority made a correct judgment. Thus, when they agreed they were right in 70% of the cases, wrong in 30%.

Practically all of the reasons given by the judges for classifying a protocol as handicapped or nonhandicapped involved psychodynamic interpretations. Some typical reasons are as follows:

Treatment of end gables of house. Anxiety reflected in shading on trees.

Although I like the person I do feel there is some distortion.

Distorted perspective on house. Anxiety reflected in tree.

Much anxiety in human drawing. Great difficulty in form.

Faltering lines. Hands omitted.

Small person in relation to other figure.

The handling of the branch structure of the tree indicates some confusion; the general quality of the person is inferior.

Insecure house. Height relation to width. Chopped off trunk which implies trauma.

Disability suggested in figure drawing.

To determine whether or not some aspect of the drawings other than the projective characteristics may have influenced the judges, the

TABLE 2

DISTRIBUTION OF MAJORITY JUDGMENT BY 9 JUDGES OF 36 MATCHED PAIRS ON RIGHT-WRONG AND GOOD-ENOUGH MENTAL AGE VARIABLES

Goodenough MA	Majority Judgment		
	Right	Wrong	Total
Handicapped lower than control	21	2	23
Handicapped higher than control	2	11	13
Total	23	13	36

Note.—One of the matched pairs of children had equal Goodenough mental ages. Tables 2 and 3 do not include this pair.

Goodenough scores were determined by an analysis of the drawing of the Person. Table 2 shows the relation between rightness or wrongness of the majority judgment and the variable of whether the handicapped child's Goodenough MA was higher or lower than that of the control.

From Table 2 it is apparent that the majority judgment was correct in 23 of the 36 possible cases, and that the Goodenough MA of the handicapped child was lower in 23 out of 36 cases, not significantly different from chance. (The average Goodenough MA difference was not significant at the .05 level.) It is evident also that there was a marked tendency for the majority judgment to be right when the Goodenough MA of the handicapped child was lower than that of the control. The majority judgment was usually wrong, however, when the Goodenough MA of the handicapped member of the pair was higher than that of his control. The chi square for the data in Table 2 was significant beyond the .001 level, after correction for continuity according to Yates. These tendencies are accentuated if only those cases were chosen where there was over two-thirds agreement among the judges. Table 3 shows the distribution of 26 "substantial agreement" cases.

Table 3 shows strikingly what has been the pattern of choices by the psychologist judges. In every case in the table, the judges (as a group) chose as the handicapped child the one with the lower Goodenough MA. In 18 out of 26 cases they were right. (It will be remembered that the average Goodenough MA of the controls was greater than that of the handicapped.) They were wrong in the eight

TABLE 3

DISTRIBUTION OF "SUBSTANTIAL AGREEMENT" JUDGMENTS ON RIGHT-WRONG AND GOODENOUGH MENTAL AGE VARIABLES

Goodenough MA	Group Judgment	
	Right	Wrong
Handicapped lower than control	18	0
Handicapped higher than control	0	8

cases where the handicapped child's Goodenough MA was *higher* than the control's.

Several of the judges gave as reasons for their decisions some evidence of self-depreciation, of a depressed self-concept in the protocols of some youngsters whom they therefore judged to be handicapped. A tendency to minify the Person was considered by some judges to be a projection of inferiority feelings arising as a result of the handicap. The height of the Person was measured as a check on this assumption. In 22 out of 37 cases, the Person drawn by the handicapped child was *larger* than that of the matched control child. This proportion is not significantly different from chance for $N = 37$.

DISCUSSION

The most likely inference is that the judges, while verbalizing psychodynamic bases for their decisions, were using primarily the intellectual characteristics of the drawings in judging which drawing belonged to the handicapped child and which to the control. They attributed the "better" drawing to the control and were right oftener than wrong, possibly because the Goodenough scores of the controls were higher in 62% of the cases than those of the handicapped children. If one took only the Goodenough MA scores and "judged" the better drawing to be that of the control, his percentage of correct judgments would be almost exactly the same as that of the total (or average) for the nine judges. It is possible, of course, that there are dynamic variables associated with "goodness" of the drawings, and the judges used these as criteria for their decisions. It is also possible that psychologists with intensive experience in the interpretation of drawings would have had a

larger percentage of correct judgments, although the study of Schmidt and McGowan (1959) suggests otherwise.

SUMMARY

H-T-P drawings of 37 pairs of elementary school age children, one physically handicapped and one a control matched for age, sex, and IQ, were judged individually by nine psychologists to postdict which drawing was made by the handicapped child. The percentage of correct judgments was not significantly above chance for any one judge, although when the judgments were pooled the combined percentage correct was significantly greater than chance. There was a strong and significant tendency for the judges to attribute the drawing with the higher Goodenough score to the control subject. The handicapped child's drawing of a Person tended to be larger than that of the control, but the difference was not significant.

REFERENCES

- BERREMAN, J. V. Implications of research in the social psychology of physical disability. *Except. Child.*, 1954, 20, 347-350, 356-357.
- BUCK, J. N. The H-T-P technique: A qualitative and quantitative scoring manual. *J. clin. Psychol.*, 1948, Monogr. Suppl. No. 5.
- CRUICKSHANK, W. M. Psychological considerations with crippled children. In W. M. Cruickshank (Ed.), *Psychology of exceptional children and youth*. Englewood Cliffs, N. J.: Prentice-Hall, 1955.
- FORCE, D. G., JR. Social status of physically handicapped children. *Except. Child.*, 1956, 23, 104-107, 132-133.
- GOODENOUGH, FLORENCE. *Measurement of intelligence by drawings*. New York: World Book, 1926.
- LENER, R., & MARTIN, M. What happens to the college student with a physical handicap? *Personnel guid. J.*, 1955, 34, 80-85.
- LEVY, J., & MICHELSON, BARBARA. Emotional problems of physically handicapped. *Except. Child.*, 1952, 18, 200-206.
- SCHMIDT, L. D., & MCGOWAN, J. F. The differentiation of human figure drawings. *J. consult. Psychol.*, 1959, 23, 129-133.
- SIERE, MARIE. Socio-emotional factors in families of the twin with cerebral palsy. *Except. Child.*, 1956, 22, 197-199, 206-208.
- WAWRZASZEK, F., JOHNSON, O. G., & SCIERA, J. L. A comparison of H-T-P responses of handicapped and nonhandicapped children. *J. clin. Psychol.*, 1958, 14, 160-162.

- WENAR, C. The effects of a motor handicap on personality: III. The effects on certain fantasies and adjustive techniques. *Child Develpm.*, 1956, 27, 9-15.
- WENAR, C. The degree of psychological disturbance in handicapped youth. *Except. Child.*, 1958, 25, 7-10.
- WHITEHOUSE, F. A. Habilitation: Concept and process. *J. Rehabil.*, 1953, 19, 3-7.
- WRIGHTSTONE, J. W. Studies of orthopedically handicapped pupils. *Except. Child.*, 1957, 23, 160-164, 176-177.

(Received May 23, 1960)

CHANGES IN INTELLECTUAL FUNCTIONS OF CHILDREN IN A PSYCHIATRIC HOSPITAL

E. WESLEY HILER AND DAVID NESVIG

Mental Health Research Institute, Fort Steilacoom, Washington

It is a well-known fact that emotional disturbances can be intellectually incapacitating, interfering with concentration, learning, memory, judgment, and reasoning. Hence it is to be expected that such disturbances will not only interfere with academic functioning, but will also impair performance on psychological tests.

Despite the fact that IQ scores are widely interpreted as measures of intellectual capacity, the experienced clinician usually regards the test scores of seriously disturbed children as measures of intellectual functioning at the time of testing rather than as representing actual intellectual potential. The latter basic capacity or potential is usually inferred from those aspects of the test performance, such as the vocabulary level, which are assumed to be less influenced by emotional disturbance. Then, too, when a patient whose general test performance is poor or mediocre does well on some of the difficult items, one is led to suspect that the actual intellectual capacity is greater than the IQ score would suggest. Marked discrepancies among the Wechsler-Bellevue subtest scores are often the basis for inferring intellectual impairment of either functional or organic origins (Wechsler, 1958).

Although the average IQ of a group of normal children usually remains constant (Brown, 1950; Gehman & Matyas, 1956), certain individuals show a marked improvement and others a marked decline in test performance during the course of childhood. Such variations have been found to be related to emotional adjustment (Allen & Young, 1943; Clarke & Clarke, 1953; Despert & Pierce, 1946). It has also been reported that the IQ often rises as a consequence of successful psychotherapy (Chidester, 1934; Dulskey, 1942;

Harrower, 1958; Hunsley, 1939; Miller, 1933) or other forms of treatment (Fisher, 1949; Markwell, Wheeler, & Kitzinger, 1953; Rabin, 1944). Change to a better environment also often leads to an improvement in test performance (Skeels & Harms, 1948). Children in warm, democratic homes were found to improve in IQ during childhood, whereas children in actively hostile and passive-neglectful homes tended to decline in IQ (Baldwin, Kalhorn, & Breese, 1945). The improvement in test performance is usually not uniform throughout the test. Certain aspects of test performance improve more than other aspects. Thus Harrower (1958) reports that a rise in Comprehension and Similarities scores on the Wechsler-Bellevue is related to clinical improvement in a group of adults. Kessler (1947) found that Picture Completion and Comprehension showed a significant rise after electroshock treatment, and Fisher (1949) reports significant improvement in Comprehension, Similarities, and Digit Symbol after EST.

The present study was carried out to investigate the effect on intellectual functions of the treatment program for children at Western State Hospital. This treatment program includes care by attendants selected for their ability to relate to children with warmth and with consistent discipline, classes conducted by teachers specially trained to deal with the emotionally disturbed, and participation in various recreational activities, arts, and crafts. An attempt is made to create a stable, noncompetitive environment with a minimum of stress. In addition, some patients receive tranquilizers and a few receive psychotherapy. Removal from an emotionally disturbing home environment and placement

in a relatively stable environment may be considered therapeutic in itself.

These children's principal intellectual deficiency was in Verbal IQ, which averaged about 14 points below Performance IQ. Therefore, we hypothesized that they would improve primarily on Verbal subtests and Verbal IQ. Inhibited, compulsively achieving children often have a Verbal IQ above the Performance IQ. In such children, one might expect Performance IQ to rise with clinical improvement. The problems of hospitalized children, however, differ from those of the typical neurotic patients seen in child guidance clinics. The child in a mental hospital is more apt to have a history of delinquent act-out and school failure. His poorly controlled hostile impulses impede his learning in school, especially his learning of verbal material. He fails to acquire the normal fund of factual information, arithmetic skill, common sense, and reading ability which directly or indirectly is measured by the Verbal section of the Wechsler. Children with these characteristics are often sent to correctional schools. Several studies have reported the Verbal IQs of delinquents as below their Performance IQs (Bernstein & Corsini, 1953; Wechsler, 1958). A similar pattern was found for un-1958). A similar pattern was found for un-1958). A similar pattern was found for un-1958). A similar pattern was found for un-1958). The more successfully disturbed delinquent child is frequently sent to a state hospital. Many of the children in this hospital are either transferred to it from correctional schools or sent to the hospital as an alternative to correctional school.

METHOD

As part of another study, all children and adolescents up to the age of 18 who were admitted to Western State Hospital after a certain date were administered a battery of psychological tests at regular intervals. Out of the group of 40 cases which had been retested, we selected all those: (a) who had been given the Wechsler-Bellevue Form II upon admission, between 2 and 3 months after admission, and between 12 and 24 months after that; (b) who had not been out of the hospital more than 6 months between the second and third testing; (c) who had attended the hospital school. This left us with a sample of 20 cases. The average age of the sample on admission was 13.8 years with a range of 10.0 to 17.8 years. Diagnostic categories included three schizophrenics, three organics with behavior disorder, nine psychoneurotics, and five character disorder.

Two additional cases, which had not had the first retest, were added to our sample for our comparisons of test improvement with rated improvement. Most of the cases in our sample had also been given the Bender-Gestalt test, from which was obtained a Pascal-Suttell Z score, the Goodenough Draw-A-Man, and the Gray Oral Reading Paragraphs Test.

On the Wechsler-Bellevue, in addition to making comparisons of IQs and subtest scores, subtests were grouped on the basis of Cohen's (1957, 1959) factor analysis of the Wechsler intelligence tests. Scores were obtained on four factors he found for the age level of our sample. The Verbal Comprehension Factor is the average of Information, Comprehension, Similarities, and Vocabulary; Perceptual Organization is the average of Object Assembly and Block Design; Freedom from Distractibility is the average of Arithmetic and Digit Span; and Judgment (Cohen's Verbal Comprehension II) is the average of Comprehension and Picture Completion.

In this study, it is assumed that changes in test scores after 1 or 2 months reflect practice effects or adjustment to the testing situation and to the hospital environment in general. The changes in scores after 12-24 months are assumed to reflect more basic changes in intellectual functioning.

Measures of behavioral changes were obtained in order to determine whether improvement in test performance was accompanied by actual improvement in condition. Change in condition was measured by a rating scale consisting of 20 variables, each on a five-point scale.

An overall rating of improvement was obtained on each patient by averaging the ratings of the specific traits. This procedure would only be justified if there were considerable homogeneity among the traits rated. A homogeneity coefficient was, therefore, computed by dividing the average between-trait (within subject) variance by the total variance, subtracting this from 1, and taking the square root. The coefficient was found to be .68; it indicates a moderate amount of homogeneity—sufficient to justify averaging the ratings on the individual traits to form a composite overall improvement score.

Each child was rated by staff members who were well acquainted with him. Altogether 25 raters were used, including 14 ward attendants, 5 school teachers, 4 social workers, 1 physician, and 1 EEG technician. The average number of raters for each child was 7.

An interrater reliability coefficient was computed by dividing the average between-rater (within subject) variance by the total variance, subtracting this from 1, and taking the square root. The coefficient was found to be .52. This is not very high. However, the use of many raters tends to counteract the unreliability of the individual raters and thus the mean ratings on each subject are believed to be sufficiently reliable to serve as measures of clinical improvement.

Different staff members rated different children; some tended to rate generally high, while others gave generally low ratings. Therefore, we measured the bias of each rater by comparing his overall-improvement

TABLE 1
CHANGES IN TEST SCORES ON RETEST

Variable	M_1	M_2	M_3	M_2-M_1	t	M_3-M_2	t
Full Scale IQ	86.70	90.15	89.70	3.45	2.46*	-.45	.27
Verbal IQ	81.15	81.50	82.45	.35	.24	.95	.65
Performance IQ	94.90	100.20	98.80	5.30	2.77*	-1.40	.61
Performance-Verbal IQ	13.75	18.70	16.35	4.95	2.27*	-2.35	.98
Subtest Total	79.45	83.30	88.00	3.85	2.20*	4.70	2.23*
Subtest AD	1.99	2.25	2.35	.26	2.87**	.10	.58
Information	4.70	4.80	5.55	.10	.11	.75	3.45**
Comprehension	6.60	6.10	7.40	-.50	1.00	1.30	3.83**
Digit Span	6.30	6.25	6.90	-.05	.08	.65	1.55
Arithmetic	4.40	4.55	5.10	.15	.44	.55	1.59
Similarities	6.85	7.55	7.35	.70	2.50*	-.20	.43
Vocabulary	6.95	6.85	6.75	-.10	.32	-.10	.32
Picture Arrangement	9.75	9.45	9.85	-.30	.52	.40	.77
Picture Completion	8.30	8.45	9.50	.15	.42	1.05	3.62**
Block Design	8.05	9.10	9.50	1.05	2.50*	.40	.84
Object Assembly	10.75	12.40	12.50	1.65	2.56*	.10	.17
Digit Symbol	6.80	7.80	7.60	1.00	2.55*	-.20	.30
Verbal Comprehension	6.25	6.33	6.75	.08	.41	.42	2.17*
Perceptual Organization	9.40	10.75	11.00	1.35	3.35**	.25	.56
Freedom from Distractibility	5.35	5.40	6.00	.05	.15	.60	1.92
Judgment	7.45	7.27	8.45	-.18	.47	1.18	5.23**
Bender Gestalt Z	99.55	95.60	84.40	-3.95	.73	-11.20	2.32*

Note.— M_1 is mean score on initial testing. M_2 is mean score on retest 2-3 months after first testing. M_3 is mean score on retest 12-24 months after second testing.

* $p < .05$.

** $p < .01$.

ment rating for each child with the overall-improvement rating for that child and taking the average of his deviations from the mean. We obtained the average rater bias for each child by averaging the bias of each of his raters. We then obtained corrected ratings for each child by subtracting the average rater bias from each child's average rating.

Our sample was small and not normally distributed on the psychological test variables; therefore, we dichotomized the scores on each variable at the median and used nonparametric statistics to compare test improvement with rated improvement. The significance level was evaluated with Fisher's exact test. The degree of relationship is indicated by the phi coefficient.

RESULTS AND DISCUSSION

Test Improvement of Group as a Whole

Table 1 contains the mean scores on the test variables on initial testing, retesting 2-3 months later, and retesting 12-24 months after that. It will be noted that on the first retest there is a rise of more than 5 points ($p = .05$) on the Performance IQ. This rise in Performance IQ results in a rise ($p = .05$) in the Full Scale IQ as well. There is no evidence of a rise in Verbal IQ. These results are consistent with other studies of practice

effects (Derner, Aborn, & Canter, 1950; Hamister, 1949; Hays & Schneider, 1951; Steisel, 1951). Because of the increase in Performance IQ and the lack of increase in Verbal IQ the already large difference between Verbal and Performance IQ in this group becomes even larger ($p = .05$).

The following subtests show a significant improvement on the first retest: Object Assembly ($p = .05$), Block Design ($p = .05$), Digit Symbol ($p = .05$), and Similarities ($p = .05$). The only factor to show a rise on the first retest is Perceptual Organization ($p = .01$). These changes are undoubtedly due at least in part to practice effect. Object Assembly, Block Design, Digit Symbol, and Picture Arrangement were reported as showing the greatest amount of practice effect after 1 and 4 weeks for a group of normals (Derner et al., 1950). Some of the improvement may reflect the stabilizing effect that the hospital has on these patients during the first few months. This stabilization would reduce loose associations and result in an improvement in the ability to perceive relationships as measured by the Similarities subtest.

Because of the marked increase on certain subtests which are already high for this group, there is an increase in subtest variability ($p = .01$). This suggests that one should be cautious in making inferences of pathology on the basis of the subtest variability of patients who have been tested before. Because practice has a greater effect on the Performance IQ than on the Verbal IQ, one would usually expect to find Performance IQ higher than Verbal IQ for individuals who have been tested previously.

After a period of 12–24 months there is an appreciable rise in the subtest total ($p = .05$), but this does not result in an improvement in IQ because by that time the children fall in a different age bracket and require a higher subtest total to achieve the same IQ. After 12–24 months the greatest rise occurs on the Information, Comprehension, and Picture Completion subtests ($p = .01$).

The scores on two of Cohen's factors show a significant rise after 12–24 months. The improvement in the Judgment factor is highly significant ($p = .01$). The improvement on the Verbal Comprehension factor was smaller ($p = .05$). A small rise on the Freedom from Distractibility factor approaches significance.

It is interesting to note that the subtests and factors showing the most improvement after 12–24 months were not the ones showing improvement after 2–3 months.

Test Improvement Related to Ratings of Improvement

There is a significant relationship ($\Phi = .55$, $p = .05$) between improvement in Verbal IQ and ratings of general improvement. Children showing more than the median amount of improvement went up 4.8 points in Verbal IQ, whereas children who showed less than the median amount went down 1.7 points.

The relationship between rated improvement and Full Scale IQ was smaller but was also significant ($\Phi = .36$, $p = .05$).

There was no significant relationship between clinical improvement and improvement in Performance IQ ($\Phi = .10$, $p = ns$).

Improvement on only one subtest, Digit Span, was significantly related to overall clinical improvement ($\Phi = .46$, $p = .05$). Information and Comprehension approached significance ($p = ns$).

Changes on the Goodenough IQ and the Bender-Gestalt did not seem to be related to clinical improvement. Improvement on the Gray reading test was, however, related to rated improvement ($N = 12$, $\Phi = .66$, $p = .05$). Those rated as improving more than the median amount went up 1.5 years in reading level while those improving less than the median amount went up only .6 years.

A comparison of each of the 20 rated variables and the psychological test variables showed the following significant relationships ($p = .05$ unless specified):

1. Improvement in Verbal IQ was related to improvement in most of the 20 clinical variables but only the following relationships were statistically significant beyond the .05 level of confidence: Achievement in School, Development of New Interests and Goals, Ability to Concentrate and Resist Distractions, Reduction in Anxiety.

2. Achievement in School and Development of New Interests and Goals were significantly related to improvement on the Information and Comprehension subtests. Reduction in Anxiety was significantly related to improvement in Digit Span. Improvement in Dependability was related to improvement on the Arithmetic subtest ($\Phi = .55$, $p = .05$). Improvement in Ability to Concentrate and Resist Distractions was related to improvement in Digit Span and the Gray reading test ($\Phi = .56$, $p = .05$).

3. Improvement on the Digit Symbol subtest was significantly related to Decrease in Bizarre Thought Processes and also to Decrease in Delinquent Tendencies.

SUMMARY

This study was carried out to determine the effect of a hospital program on the intellectual functioning of emotionally disturbed children.

A sample of 20 children with a mean age of 13.8 years was tested on admission to the hospital, retested 2–3 months later, and retested again 12–24 months after that.

A significant rise occurred after 2–3 months on Wechsler Performance IQ, Full Scale IQ, subtest total, subtest average deviation, Similarities, Block Design, Object Assembly, and Digit Symbol, and the Perceptual Organization factor. This improvement is partially at-

tributable to practice effect, but the improvement in Similarities perhaps reflects a decrease in bizarre or irrelevant thought processes.

After 12-24 months, a marked improvement was shown on the Information, Comprehension, and Picture Completion subtests; the Judgment and Verbal Comprehension Factors; and Bender-Gestalt performance. The group as a whole appears to be better organized perceptually, to have more common sense, better judgment, and an increased ability to perceive relationships and distinguish between essential and unessential aspects of a situation.

The Performance, Verbal, and Full Scale IQs did not go up for the group as a whole. However, it was noted that certain patients did improve considerably on these scales while others declined. It was hypothesized that such differences in the direction of change would be related to improvement or deterioration of the patient's condition.

Ratings of improvement on 20 clinical variables were obtained from the hospital staff. Improvement in Verbal IQ, Full Scale IQ, Digit Span, and Gray Oral Reading test level was significantly related to overall clinical improvement. Improvement in IQ and in specific subtests was found to be related to improvements on specific traits.

It is concluded that most hospitalized children have problems which cause them to be retarded in verbal skills; as they improve, their Verbal IQ rises. It is suggested that the initial Verbal IQ is not a fair estimate of the intellectual capacities of emotionally disturbed children, and that Performance IQ may provide a more accurate measure of these children's actual intelligence.

REFERENCES

- ALLEN, M. E., & YOUNG, F. M. The constancy of the intelligence quotient as indicated by retests of 130 children. *J. appl. Psychol.*, 1943, 27, 41-60.
- BALDWIN, A. L., KALHORN, J., & BREESE, F. H. Patterns of parent behavior. *Psychol. Monogr.*, 1945, 58(3, Whole No. 268).
- BERNSTEIN, R., & CORSINI, R. Wechsler-Bellevue patterns of female delinquents. *J. clin. Psychol.*, 1953, 9, 176-179.
- BROWN, G. L. On the constancy of the IQ. *J. educ. Res.*, 1950, 44, 151-153.
- CHIDESTER, L. Therapeutic results with mentally retarded children. *Amer. J. Orthopsychiat.*, 1934, 4, 464-472.
- CLARKE, A. D. B., & CLARKE, A. M. How constant is the IQ? *Lancet*, 1953, 265, 877-880.
- COHEN, J. The factorial structure of the WAIS between early adulthood and old age. *J. consult. Psychol.*, 1957, 21, 283-290.
- COHEN, J. The factorial structure of the WISC at ages 7-6, 10-6, and 13-6. *J. consult. Psychol.*, 1959, 23, 285-299.
- DERNER, G. F., ABORN, M., & CANTER, A. H. The reliability of the Wechsler-Bellevue subtests and scales. *J. consult. Psychol.*, 1950, 14, 172-179.
- DESPERT, J. L., & PIERCE, H. O. The relation of emotional adjustment to intellectual function. *Genet. psychol. Monogr.*, 1946, 34, 3-56.
- DULSKY, S. G. Affect and intellect: An experimental study. *J. gen. Psychol.*, 1942, 27, 199-220.
- FISHER, K. A. Changes in test performance of ambulatory depressed patients undergoing electroshock therapy. *J. gen. Psychol.*, 1949, 41, 195-232.
- GEHMAN, I. A. H., & MATYAS, B. P. Stability of the WISC and Binet tests. *J. consult. Psychol.*, 1956, 20, 150-152.
- GRAHAM, E. E. Wechsler-Bellevue and WISC scattergrams of unsuccessful readers. *J. consult. Psychol.*, 1952, 16, 268-271.
- HAMISTER, R. The test-retest reliability of the Wechsler-Bellevue intelligence test (Form I) for a neuropsychiatric population. *J. consult. Psychol.*, 1949, 13, 39-43.
- HARROWER, MOLLY. *Personality change and development as measured by projective techniques*. New York: Grune & Stratton, 1958.
- HAYS, W., & SCHNEIDER, B. A test-retest evaluation of the Wechsler Forms I and II with mental defectives. *J. clin. Psychol.*, 1951, 7, 140-143.
- HUNSLEY, Y. L. Intelligence, as reflected by work habits, attitude and behavior, does change. *Sch. Soc.*, 1939, 50, 682-684.
- KESSLER, LUCILLE. Intellectual changes in schizophrenic patients following electroshock therapy. Unpublished master's thesis, New York University, 1947.
- MARKWELL, E. D., JR., WHEELER, W. M., & KITZINGER, HELEN. Changes in Wechsler-Bellevue test performance following prefrontal lobotomy. *J. consult. Psychol.*, 1953, 17, 229-231.
- MILLER, E. Emotional factors in intellectual retardation. *J. ment. Sci.*, 1933, 79, 614-625.
- RABIN, A. I. Fluctuations in the mental level of schizophrenic patients. *Psychiat. Quart.*, 1944, 18, 78-92.
- SKEELS, H. M., & HARMS, IRENE. Children with inferior social histories: Their mental development in adoptive homes. *J. genet. Psychol.*, 1948, 72, 283-294.
- STEISEL, I. M. The relation between test and retest scores on the Wechsler-Bellevue scale (Form I) for selected college students. *J. genet. Psychol.*, 1951, 79, 155-162.
- WECHSLER, D. *The measurement and appraisal of adult intelligence*. Baltimore: Williams & Wilkins, 1958.

(Received June 6, 1960)

A COMPARISON OF SOCIAL AND SOLITARY MALE DELINQUENTS¹

MARY H. RANDOLPH, HAROLD RICHARDSON, AND RONALD C. JOHNSON

San Jose State College

Most contemporary theories concerning the causes of juvenile delinquency might be roughly differentiated into two areas of emphasis: psychologically oriented theories, such as those expressed by the Gluecks (1950), Healy and Bronner (1936), and Lindner (1944); and sociologically oriented theories, such as those expressed by Shaw and McKay (1942), Sutherland (1955), Thrasher (1936), and others. Both points of view may be defended since it seems likely that psychological and sociological forces interact to varying degrees in the histories of most delinquents.

Lindesmith and Dunham (1941) have differentiated between the socialized and the individualized criminal. They state that the socialized criminal is one who commits crimes that are supported and prescribed by his culture, so that, by committing a crime, the criminal gains in status and recognition. The socialized delinquent or criminal acts in close collaboration with other persons and is dependent on them for the continuation of his criminal career. The individualized criminal, on the other hand, acts for reasons that are personal and private. He commits his crimes alone and, in theory, is a stranger to others who commit similar crimes. His criminal act is not an acceptable form of behavior in his social milieu. The socialized criminal seems likely to be a rather normal person, in a psychological sense, who holds deviant social values common to his subcultural group. The individualized criminal, at odds with his own primary groups, seems likely to be an individual whose criminality is merely symptomatic of deeper psychological pressures. Johnson (1949) has suggested that the solitary delinquent is an individual with a "conscience

defect" unconsciously fostered by the parents, while the social delinquent is the product of a subculture with delinquent values. Bloch and Flynn (1956) made similar statements. Within this framework, the sociological theories would seem most useful in explaining the delinquency of male social delinquents—gang members and others who commit delinquent acts in the company of others, while psychological explanations might best account for the individualized or solitary delinquent.

It has been found (Hewitt & Jenkins as reported by Wattenberg & Balistrieri, 1950) that juvenile gang members are likely to come from homes of a lower socioeconomic stratum, while nongang members showed more indications of coming from stressful or depriving homes of a middle socioeconomic level. Johnson (1950) found that solitary delinquents were far more often recidivists than were social delinquents, even though the majority of his social delinquents were members of well organized juvenile gangs. This finding might be expected if individual or solitary delinquents are merely acting out symptoms of deep-seated and unresolved psychological stresses.

Beyond these scanty data, little is known about genotypic or phenotypic variation between solitary and social delinquents, even though treatment techniques used for the two groups might be made more effective if this information were available. The purpose of this study is to compare solitary and social delinquents with regard to several "sociological" and "psychological" variables.

METHOD

Subjects

The sample consisted of 62 boys, aged 14-18, who had been adjudged legally as juvenile delinquents. Of these, 52 subjects were at a "ranch" for delinquent boys and the other 10 were in custody, awaiting

¹ This report is based on a thesis submitted (by the first author) to the Department of Psychology, San Jose State College, January 1960.

TABLE 1

NUMBER OF SOLITARY AND SOCIAL DELINQUENTS FROM EACH SOCIOECONOMIC LEVEL

Delinquent	Upper Middle	Lower Middle	Upper Lower	Lower Lower	N
Solitary	4	7	5	2	18
Social	1	4	15	19	39

Note.— $\chi^2=15.83$, $p < .01$.

ing placement at this ranch. All boys were of at least dull normal intelligence. Of the original sample, one subject was eliminated because of insufficient responses to test items, and four subjects were eliminated because extensive further examination showed them to have had mixed (solitary and social) delinquent careers. Fifty-seven subjects remained. Of these subjects, 39 had always been social and 18 had always been solitary in their known delinquencies.

Measuring Devices

Each subject was administered a Wechsler Adult Intelligence Scale (WAIS), and a Minnesota Multiphasic Personality Inventory (MMPI). Socioeconomic status was determined by a local adaptation (Hodges, unpublished) of the Warner Index (Warner, 1949).

Procedure

All subjects were tested inside an institutional setting. Tests were administered and scored according to standardized procedure except that all questions on the MMPI were read aloud to subjects, while the subjects read the questions in the booklet, in order to minimize difficulties in comprehension. All subjects knew that test results were confidential and would not influence placement. Only two MMPI records (both of social delinquents) had to be discarded as invalid.

RESULTS

On the WAIS IQ scores the social delinquents had a mean of 93.23 with a standard deviation of 9.14. The solitary delinquents had a mean of 105.00 with a standard deviation of 11.19. The t test of differences between these means was significant beyond the .01 level. The F test of differences between variances was not significant. The mean IQ score of the solitary delinquents was exceeded by only 15% of the social delinquents.

Many more solitary delinquents came from upper socioeconomic levels than did social delinquents, as shown in Table 1.

MMPI profiles are presented in Figure 1.

Mean differences between social and solitary delinquents on the validating scales L ,

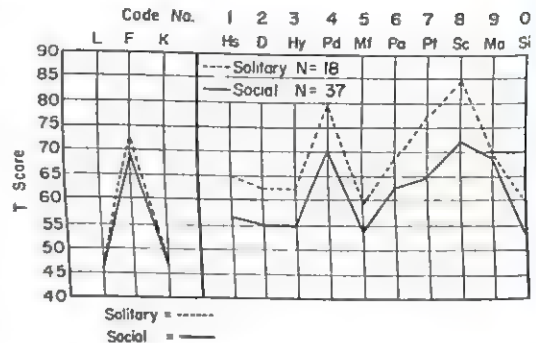


FIG. 1. Mean profiles for social and solitary delinquents on the MMPI.

F , and K were not significant. Both groups scored rather high on F , found to be an indicator of psychopathology (Kazan & Scheinberg, 1945; Modlin, 1956) and of delinquency (Hathaway & Monachesi, 1953). Profiles of the two groups are similar but solitary delinquents, as a group, appear somewhat more disturbed. All mean differences for the diagnostic scales were significant except for the Ma scale. Differences on Mf , Pa , and Si were significant beyond the .05 level. Differences on Hs , D , Hy , Pd , Pt , and Sc were significant beyond the .01 level of confidence. The code type of the solitary delinquent was 8479'612305-, (4, 13, 11). The social delinquent's code type was '8497613-, (3, 11,

TABLE 2

DATA COMPARING SOCIAL AND SOLITARY DELINQUENTS ON THE MMPI

(Social $N = 37$, Solitary $N = 18$)

Scale	Mean Social	Mean Solitary	SD Social	SD Solitary	t
L	3.45	3.55	2.42	2.04	1.51
F	11.23	12.76	4.58	4.55	1.49
K	10.15	10.94	3.97	6.62	.68
Hs	14.03	17.05	3.10	4.03	2.97**
D	18.77	21.94	3.98	4.71	3.23**
Hy	19.21	23.33	4.35	4.47	3.10**
Pd	27.62	30.89	4.38	4.63	3.05**
Mf	22.18	25.38	4.18	4.71	2.34*
Pa	12.36	14.38	2.95	3.27	2.22*
Pt	30.26	35.55	5.95	6.29	3.37**
Sc	33.41	39.83	7.61	7.93	3.54**
Ma	24.18	25.05	4.17	4.70	.78
Si	28.85	32.55	6.09	8.27	2.48*

Note.—These figures are expressed in MMPI raw scores.
* Significant at .05 level.
** Significant at .01 level.

10). Although both types had high excitors (*Pd*, *Sc*, *Ma*), only the solitaires had high suppressors (*Si*, *D*, *Mf*), presumably indicating neurotic trends. Complete statistical data for comparing social and solitary delinquents on the MMPI are given in Table 2.

DISCUSSION

It seems relatively clear, from these results, that solitary and social delinquents differ considerably from each other. The solitary delinquent seems likely to come from a higher socioeconomic level and to be of higher intellectual ability than the social delinquent, but to be considerably more maladjusted. These findings might explain why, once embarked upon a course of delinquent behavior, the solitary delinquent is more inclined to be a recidivist.

Differences between the two groups of delinquents might be taken into account in diagnosis, prognosis, and treatment. The prognosis for the solitary delinquent without some form of therapy seems likely to be poorer than for social delinquents. Current sociologically oriented treatment techniques might more often be sufficient for the rehabilitation of the social delinquent. So long as therapy services are in short supply, a more economical use of therapists' time might result from a consideration of the obtained differences between social and solitary delinquents.

SUMMARY

Test data were obtained from 57 delinquent boys, aged 14-18, all of whom, at the time of testing, were within an institutional setting. Thirty-nine of the subjects were "social delinquents" who committed their crimes in the company of others. Eighteen subjects were "solitary delinquents" who had committed their delinquencies alone. Wechsler intelligence test (WAIS) scores indicated that solitary delinquents are significantly higher than social delinquents in intellectual ability. Solitary delinquents were also significantly higher in socioeconomic status than were social delinquents. MMPI profiles of the two groups were similar, but with a significantly greater elevation in all scales except *Ma* for the soli-

tary delinquent. The solitary delinquent appears more likely to be a psychologically deviant individual who comes from an ostensibly normal environment, while the social delinquent seems far less deviant, in a psychological sense, but comes from an environment where certain sociological factors, presumably causal to delinquency, are operating. Certain implications of these findings were discussed.

REFERENCES

- BLOCH, H. A., & FLYNN, F. T. *Delinquency*. New York: Random House, 1956.
- GLUECK, S., & GLUECK, ELEANOR T. *Unraveling juvenile delinquency*. New York: Commonwealth Fund, 1950.
- HATHAWAY, S. R., & MONACHESE, E. D. *Analyzing and predicting juvenile delinquency with the MMPI*. Minneapolis: Univer. Minnesota Press, 1953.
- HEALY, W., & BRONNER, AUGUSTA F. *New light on delinquency and its treatment*. New Haven: Yale Univer. Press, 1936.
- JOHNSON, ADELAIDE M. Juvenile delinquency. In S. Arieti (Ed.), *American handbook of psychiatry*. Vol. 1. New York: Basic Books, 1949.
- JOHNSON, R. C. Causal factors in the delinquency of fifty Denver boys. Unpublished master's thesis, University of Denver, 1950.
- KAZAN, A. T., & SHEINBERG, I. M. Note on the significance of the validity score (*F*) in the MMPI. *Amer. J. Psychiat.*, 1945, 102, 181-183.
- LINDESMITH, A. R., & DUNHAM, H. W. Some principles of criminal typology. *Soc. Forces*, 1941, 19, 307-314.
- LINDNER, R. *Rebel without a cause*. New York: Grune & Stratton, 1944.
- MODLIN, H. C. A study of the MMPI in clinical practice. In G. Welsh & W. G. Dahlstrom (Eds.), *Basic readings on the MMPI in psychology and medicine*. Minneapolis: Univer. Minnesota Press, 1956.
- SHAW, C. R., & MCKAY, H. D. *Juvenile delinquency in urban areas*. Chicago: Univer. Chicago Press, 1942.
- SUTHERLAND, E. H., & CRESSEY, D. R. *Principles of criminology*. Philadelphia: Lippincott, 1955.
- THRASHER, F. M. *The gang*. Chicago: Univer. Chicago Press, 1936.
- WARNER, W. L., MEEKER, M., & EELS, K. *Social class in America*. Chicago: Science Research Associates, 1949.
- WATTENBERG, W. W., & BALISTIERI, J. J. Gang membership and juvenile misconduct. *Amer. sociol. Rev.*, 1950, 15, 744-752.

(Received June 6, 1960)

THE DIMENSIONALITY OF RATINGS OF THERAPIST VERBAL RESPONSES¹

EDMUND S. HOWE AND BENJAMIN POPE

University of Maryland School of Medicine

The last few years have witnessed an important trend in research in psychotherapy, toward study of the therapist as an independent variable in the dyadic relationship. This trend has in part shifted the focus of empirical attention away from such issues as theoretical differences per se among "schools" of psychotherapy, and has instead directed research toward rigorous quantification of basic variables cutting across theoretical and practical divergences among therapists. Among the most penetrating studies of this kind are those investigating the dimension and the dimensionality of Depth of Interpretation (e.g., Harway, Dittmann, Raush, Bordin, & Rigler, 1955; Raush, Sperber, Rigler, Williams, Harway, Bordin, Dittmann, & Hayes, 1956; Speisman, 1959). Other investigators have approached presumably different aspects of therapist verbal behavior such as Directiveness (e.g., Snyder, 1953) and Ambiguity (e.g., Osburn, 1951), to mention but two.

Howe and Pope (1961) and Pope, Howe, and Finesinger (1959) have more recently approached the problem of therapist verbal behavior from the standpoint of Finesinger's (1948) enunciation of a principle of Minimal

Activity. Using attributes of "Ambiguity," "Lead," and degree of "Inference" as aspects of the concept of Therapist "Activity," an Activity Scale was constructed on the basis of ratings, by psychiatrists, of a representative sample of 50 therapist verbal responses. The order of reliability observed among ratings used in constructing the Activity Scale and ratings obtained from *application* of the scale was about .50. While this average reliability coefficient compares quite favorably with those reported by earlier investigators of a different aspect of therapist verbal behavior (e.g., Harway et al., 1955; Raush et al., 1956) 75% of the total variance is nevertheless left as unexplained "error." The rather low reliability indices obtained in these and other studies is quite possibly due to the assumed "dimension" in rating studies being multiple, rather than unitary; as Raush et al. (1956) earlier pointed out. For, as Coombs (1951) and Bordin, Cutler, Dittmann, Harway, Raush, and Rigler (1954) have written, it is quite possible to "force" unidimensionality even where it does not mathematically exist. Indeed, even while the studies of therapist activity were being performed, it became subtly obvious to the experimenter that the resistance of some of the psychiatrists to performance of the rating task devoid of a flesh-and-blood patient was at least in part due to an apparent confounding of their *evaluative* attitudes with their judgments of activity according to relative ambiguity, lead, and inference involved in each therapist response. Were this the case, then one would statistically predict relatively low reliability among ratings along a one-dimensional continuum.

The study reported here was thus undertaken to explore the dimensionality of ratings of such types of therapist verbal responses as

¹ This paper arose out of research supported by Pilot Evaluation Grant No. 2M-6408 from the National Institute of Mental Health of the National Institutes of Health, United States Public Health Service. The late Jacob E. Finesinger was the principal investigator. Thanks are acknowledged for his continuous encouragement and wholehearted support of this work until his untimely death in June 1959. Final completion of the work and of the present manuscript was indirectly facilitated by Research Grant M-3355 (also from the National Institute of Mental Health) of which ESH is the principal investigator. A paper based upon this research was presented to Division 12 at the Annual Convention of the American Psychological Association in Chicago, September 1960.

had been used in the earlier presumptive one-dimensional studies of therapist activity. Application of a 40-scale semantic differential to therapist responses thus facilitated a check on the two general propositions that such ratings of therapist verbal responses would be primarily of an evaluative nature, and at least two-dimensional.

METHOD

Raters. A decision was made to solicit the services of Board-certified or Board-eligible psychiatrists, rather than of psychiatrists having had some minimum amount of therapeutic experience. The subjects were drawn from the Psychiatric Institute at the University of Maryland School of Medicine, from those engaged in full-time private practice in Baltimore City, from the National Institute of Mental Health, from the Walter Reed Army Hospital and Institute for Research, from Chestnut Lodge, and from Spring Grove State Hospital, Maryland. The booklets described below were mailed, after verbal agreement by telephone, to 50 subjects. Of the 38 booklets returned,² 3 were discarded because of inadvertent omission of at least 1 page, either by experimenter or by subject. Data from the remaining 35 subjects were analyzed.

Therapist Verbal Responses. Ten therapist responses from the set of 50 used in the earlier rating study were selected for experimentation. These responses, each chosen for specific reasons of either close a priori similarity to or extreme a priori dissimilarity from at least one other response, are presented in Table 1. These reasons are now briefly summarized. Responses 1 and 10 were selected for study because they were originally given the most extensive mean Activity Level (AL) ratings. Responses 2 and 3 were selected because they were originally rated equally, and were clearly of a very low-active, facilitating nature. Responses 4 and 5 were likewise chosen since both were about equally and yet somewhat more specifically focused than 2 and 3, and both still quite low-active. Response 6 is unique in the set of 10 studied here, since it is a highly specific, *objective* question. Responses 7 and 8 were included since they constitute interpretive responses of subjectively equal depth, and have equal ALs. Response 9 is evidently a reassurance/supportive type of response, subjectively quite distinctive in its own right.

Since the 10 responses could clearly be used as "markers," the potential disadvantage of making a deliberate selection (with the attendant risk of bias) was largely offset by the definitive role the responses could play in the face validation of certain empirical outcomes to be described later. The question whether Response 10 (and perhaps even Response 9) nor-

TABLE 1

THE 10 THERAPIST VERBAL RESPONSES PRESENTED TO THE RATERS AND ORIGINAL MEAN ACTIVITY LEVELS

No.	Therapist Verbal Response	AL
1	Hm-hm	1.4
2	And —?	2.4
3	What's been happening?	2.6
4	By <i>brooding</i> you mean —?	4.3
5	Your <i>heart</i> —?	4.4
6	How much do you earn?	6.2
7	Perhaps he feels attracted to you	7.8
8	Maybe you participate in this, encouraging them to depend on you more than you think.	7.8
9	I hope you don't think I'm impatient with you, because I'm not	8.7
10	You have to tell me whether it is so or not.	9.7

Note.—Based on the ratings of 30 psychiatrists during the first, 50-item rating study.

mally occur too infrequently to be included without risk of seriously biasing the empirical outcomes in this type of research can not be fully dealt with here, but it deserves comment. Response 10 ("You have to tell me whether it is so or not") is essentially a demanding and persuasive operation instigated by the patient. As such, it is largely a tabooed response, negatively valued by most therapists. (A comparable, though weaker argument might be made by some therapists with regard to Response 9, which is essentially a reassurance operation instigated by the therapist himself.) Operations of a persuasive nature may, however, take many and variegated forms. Persuasion qua persuasion rarely occurs in most accepted psychotherapies. But the type of persistence and hounding of the patient's thoughts that occurs in the published interviews of Deutsch and Murphy (1955), to cite only one example, reflects from an operational standpoint a respectable unwillingness of the therapist to let the patient "get away" until he reports that which the therapist, perhaps unconsciously, wants to hear. Thus, if the concept of persuasiveness be regarded more as an "attitude of mind" in the therapist, than as the manifest form that his communications to the patient actually take, then it becomes more reasonable to include some sort of verbal stimulus connoting such a therapist attitude. Indeed, inclusion of such an anchor stimulus would, under the foregoing conditions, be as like to *reduce* as to foster bias in the empirical outcome.

² Many thanks are here expressed to all of those unnamed persons who were kind enough to perform the ratings.

TABLE 2
THE SEVEN SETS OF BIPOLAR,
ADJECTIVAL SCALES

Hypothetical Variable	Bipolar Scale
Ambiguity	Spacious-Constricted Colorless-Colorful General-Specific Unfocused-Focused Commonplace-Unique Vague-Precise
Stressfulness	Cold-Warm Calm-Excitable Sober-Drunk Relaxed-Tense Cautious-Rash Relieving-Painful
Inference	Subtle-Obvious Inferential-Logical Intuitive-Rational Deep-Shallow Private-Public
Lead	Following-Leading Accepting-Demanding Conforming-Directing
Evaluative	Skillful-Unskillful Reputable-Disreputable Wise-Foolish Good-Bad Accepting-Rejecting Sensitive-Insensitive Acceptable-Unacceptable Valuable-Worthless
Activity	Still-Vibrant Static-Dynamic Slow-Fast Inert-Energetic Passive-Active
Potency	Muted-Blatant Weak-Strong Soft-Hard Thin-Thick Far-Near Small-Large Dull Sharp

Note.—The classification is arbitrary in several cases.

The Rating Booklet. Each rater was presented with a 20-page booklet. Each successive pair of pages contained a total of 40 seven-point, bipolar, adjectival scales, the therapist response to be judged appearing at the top of each pair of pages. The 40 scales were of course the same for each response, and

they appeared in the same order. The order in which the therapist responses appeared, however, was varied in four ways (viz.: Numbers 1-10; Numbers 10-1; Numbers 6-10, 1-5; Numbers 5-1, 10-6). In all other respects the format of the instructions to the subject followed that described by Osgood, Suci, and Tannenbaum (1957). The subject was instructed to assume that each response was made during an initial interview. It was considered desirable to permit the subject to project his own feelings about context, since the generality of the findings would thereby be enhanced.

The Adjectival Scales. The set of 40 scales was selected after an exhaustive examination both of *Rogel's Thesaurus*, and of published work (e.g., Osgood et al., 1957) using different forms of the Semantic Differential. It was decided to include scales having some connotative reference to Ambiguity, Lead, and Inference, since these attributes had been used earlier as rough working referents of the concept of Activity. Scales having connotative reference to Stressfulness were also included because of such frequent explicit claims by previous subjects as, "I would consider something more 'Active' if it tends to upset the patient." Finally, in view of their ubiquitous appearance in numerous reported studies, scales were included having established relevance to Osgood's Evaluative, Potency, and Activity dimensions. These seven sets of bipolar adjectives are presented in Table 2. The present classification of individual items into the seven sets is, of course, purely arbitrary in several instances.

Treatment of Data. A 40×40 intercorrelation matrix, with $N = 35$ (Judges) $\times 10$ (Verbal Responses) = 350 pairs entering into each correlation, was obtained by IBM. The matrix was factored by the complete centroid method of Thurstone (1947), and since a maximum of seven hypothetical dimensions was implied by the initial categorization of the scales into seven sets, a total of nine factors was extracted, of which three were significant and interpretable. The centroid was orthogonally rotated by the quartimax method (Neuhaus & Wrigley, 1954).

RESULTS

The Factor Analysis

The rotated factor loadings are presented in Table 3. The first factor to emerge accounts for 33% of the total variance and 60% of the common variance. It is most clearly defined by the bipolar scales foolish-wise, unacceptable-acceptable, skillful-unskillful, good-bad, and valuable-worthless. The first two scales have rather pure, high negative loadings and the last three rather pure, high positive loadings on the first factor. Other, somewhat less pure but still significant loadings are observed for scales tense-relaxed, blatant-muted, and rejecting-accepting (negatively

TABLE 3
ROTATED FACTOR LOADINGS

Scale	I	II	III	h^2
Tense-Relaxed	-.83	-.16	.03	.72
Skillful-Unskillful	.87	-.02	-.04	.76
Hard-Soft	-.78	-.26	-.06	.68
Passive-Active	.38	.65	-.03	.57
Near-Far	.52	-.34	.01	.39
Reputable-Disreputable	.80	-.01	-.03	.64
Spacious-Constricted	.60	.11	.13	.39
Foolish-Wise	-.90	.02	.10	.82
Cautious-Rash	.83	.27	-.03	.76
Blatant-Muted	-.83	-.28	-.08	.77
Colorless-Colorful	.09	.77	-.09	.61
Still-Vibrant	.17	.77	-.10	.63
Leading-Following	-.25	-.52	.04	.33
Small-Large	-.13	.54	-.09	.32
Accepting-Demanding	.75	.16	.20	.63
Strong-Weak	.15	-.73	-.05	.56
Specific-General	-.20	-.69	-.13	.53
Good-Bad	.90	-.04	-.05	.81
Rejecting-Accepting	-.82	.01	-.11	.68
Static-Dynamic	-.32	.60	-.17	.49
Conforming-Directing	.38	.46	.07	.36
Thick-Thin	-.11	-.48	.10	.25
Focused-Unfocused	-.17	-.63	-.06	.43
Slow-Fast	.19	.67	-.12	.50
Obvious-Subtle	-.61	-.14	-.30	.48
Sensitive-Insensitive	.84	-.10	.14	.74
Calm-Excitable	.83	.18	.00	.72
Energetic-Inert	-.16	-.82	.02	.70
Cold-Warm	-.61	.27	-.09	.45
Inferential-Logical	.16	-.17	.74	.60
Unique-Commonplace	-.08	-.38	.28	.23
Vague-Precise	.07	.71	.14	.53
Unacceptable-Acceptable	-.88	.00	.08	.78
Relieving-Painful	.71	.11	.09	.52
Sharp-Dull	-.27	-.41	.05	.24
Sober-Drunk	.51	-.21	-.14	.32
Shallow-Deep	-.61	.37	-.11	.52
Private-Public	.37	-.26	.16	.23
Intuitive-Rational	.24	-.05	.78	.67
Valuable-Worthless	.89	-.09	-.08	.81
$\Sigma a^2 = 13.38$		7.15	1.66	$\Sigma h^2 = 22.19$

loaded), and for scales cautious-rash, sensitive-insensitive and calm-excitable (positively loaded). There is no question but that the first factor may be appropriately interpreted as one of Professional Evaluation. Its nature implies that the "good" therapist, the one who is thoroughly reputable and skilled, uses responses which are cautious, relaxed, muted, accepting, sensitive, and calm. The emergence of such a factor *first*, further implies, as we shall see, that such evaluative connotations of

the verbal responses are more compelling, more salient than are the connotations of the second and subsequent factors interpreted.

The second factor accounts for 18% of the total variance and 32% of the common variance. It is most clearly defined by scales colorless-colorful, still-vibrant, and vague-precise (positive loadings), and by scales energetic-inert and strong-weak (negative loadings). Other, somewhat less pure but still significant loadings on the second factor are ob-

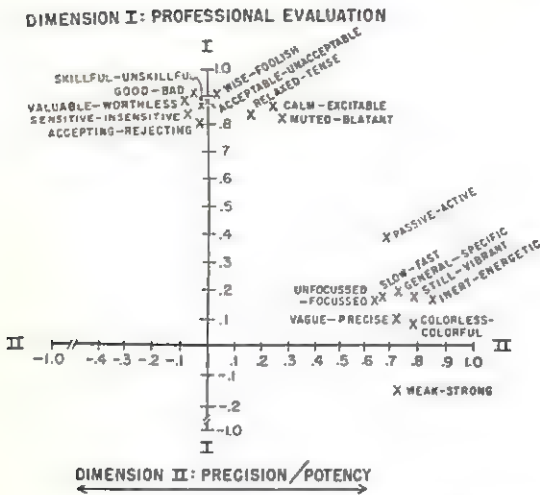


FIG. 1. Scales having significant rotated loadings on Factor 1, or on Factor 2.

served for scales passive-active and slow-fast (positive loadings) and for scales specific-general and focused-unfocused (negative loadings). The second factor is interpreted as one of Precision/Potency, although Ambiguity/Passivity would be almost as appropriate a label. This factor clearly refers to those attributes of therapist behavior variously referred to as Activity, Ambiguity, and the like. Its nature is reminiscent of the "dynamism factor" so labeled by Osgood et al. (1957) to describe the apparent coalescence of their second (Activity) and third (Potency) factors in the judgment of sociopolitical concepts.

The third factor, which accounts for only a little over 4% of the total variance and 7% of the common variance, is represented by only two scales: inferential-logical and intuitive-rational, both being quite highly positively loaded on this factor. The rotated loadings for these scales are, respectively, .74 and .78, and they are fairly pure scales; but there are no other scales even approaching significant loadings on this factor. Consequently, it is rather difficult to interpret this factor with great conviction; for it is probably more a factor of Subjectivity/Objectivity rather than one of inference. The difficulty in interpreting this factor reflects in part a selection of scales of which the referents, retrospectively considered, are somewhat nebulous. The fourth through sixth factors are all significant ($p < .05$) according to Humphreys'

Rule (Fruchter, 1954, pp. 79-80), the respective percentages of total variance accounted for being 3 for the fourth factor, 2 for the fifth, and 1% for the sixth factor. None of these factors, however, makes any interpretive sense.

Graphic Representation of Factors 1 and 2

A number of scales with high loadings on the first two factors are plotted two-dimensionally in Figure 1, for illustrative purposes. The axes in the Figure are drawn at right angles since the Quartimax rotation leads to an orthogonal solution. The first dimension, that of Professional Evaluation, in a sense sets the image that the raters have of "good" professional behavior; that is, responses involving connotations of acceptance, sensitivity, relaxedness, muteness, calmness, caution, and so on. The second dimension, Precision/Potency, is clearly more allied, as noted earlier, to the original concept of Activity, and its assumed attributes of Lead and Ambiguity.

Construction of the Three-Dimensional Model

A model representing the positions of the 10 therapist responses in "semantic space"

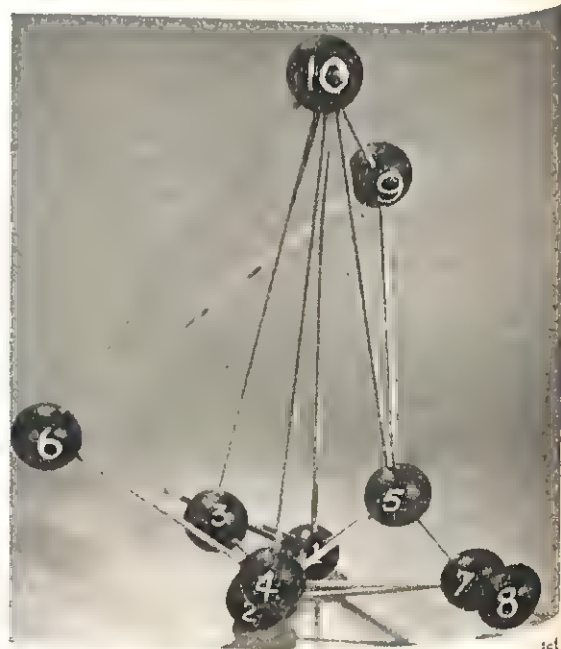


FIG. 2. Three-dimensional model of the 10 therapist responses in "semantic space".

was constructed³ from the *Generalized Distance Formula* (Osgood et al., 1957, pp. 90-97). For this purpose the raw scores on three pairs of scales were used to compute values of *D*. The three pairs of scales were selected on the basis of their having high, significant, and relatively pure loadings, respectively, on Factors 1, 2, and 3. The scales were: foolish-wise and valuable-worthless on Factor 1, colorless-colorful and energetic-inert on Factor 2, and inferential-logical and intuitive-rational on Factor 3 (see Table 3). Following Osgood's procedure, the model was represented by 10 rubber balls 1 inch in diameter, fitted together in three-dimensional space with aluminum rods of $\frac{1}{8}$ inch gauge, cut to appropriate lengths. The model is presented in Figure 2. Slight but not at all serious difficulty was experienced in fitting all of the distances exactly, thus suggesting that three dimensions account quite well for the obtained data.

Table 1 earlier presented the 10 verbal responses, and a rationale for their inclusion in the study was given. Figure 2 illustrates that, by and large, a priori expectations are borne out; and the model thus has considerable face validity. It will be recalled that Response 1 consists of "Hm-hm," while Response 10 smacks of a rare attempt to persuade the patient. These two responses are the most distant from each other in the model. Responses 2 and 3 are similar, low active facilitation responses, and they are appropriately close to each other in Figure 2. Responses 4 and 5 refer to more sharply (but equally) focused facilitating responses, and they also are quite close to each other in the model. Response 6 refers to a highly specific, objective (and presumably uncharged) question; it is unique in the set of 10, and accordingly rather solitary. Responses 7 and 8, on the other hand, constitute interpretive operations of about equally moderate depth, and are adjacent in Figure 2. The ninth response refers to a supportive/reassurance operation, clearly discriminated in the three-dimensional model from other responses. The relative positions of the 10 balls in the model thus accord largely with purely subjective clinical "feel,"

³ Thanks are acknowledged to Michael S. Black, now at the University of Illinois, for painstakingly constructing this model.

and agree very well with empirically observed groupings of the responses found in the earlier one-dimensional study.

DISCUSSION

The general findings are reminiscent of two earlier studies. One, published by Fisher (1956), concerned ratings of "plausibility" versus "depth" of interpretive responses. Fisher showed strong, significant relationships between ratings of depth of therapist interpretive operations, and ratings obtained when the working dimension was Plausibility. The findings of the present study independently suggest Fisher's results to be extremely plausible! Presumably there is a (theoretically) infinite number of "one-dimensional" scale labels that would give approximately comparable results in any such rating study.⁴ This consideration argues, of course, for the hypothesis that regardless of what *instruction* one gives to a rater, he will, in the final analysis, rate according to certain internal mediating cues which only partly correspond with whatever *explicit* cues the experimenter is trying to communicate. Fisher has drawn attention to the essence of the problem here involved. The present findings rather forcefully suggest that upon closer inspection, some of the dimensions of therapist verbal behavior frequently studied empirically may turn out, operationally speaking, to be one and the same.

A second finding of which the present ones are reminiscent was published by Raush et al. (1956). While those authors concluded that on the whole they could not find evidence of

⁴ As a matter of fact, a set of 35 abstract descriptions of therapist verbal responses were recently sorted here by completely naive, unsophisticated, freshman undergraduates, along "any increasing dimension that you think would be appropriate." The rank orderings of these therapist responses sorted by 17 subjects showed a rho of .87 with ranked mean ratings (of the same 35 responses) made by 20 Board-certified psychiatrists given a working definition of Activity Level in terms of Ambiguity, Lead, and Inference, and asked to sort on this dimension! These data are unpublished. They lead one to the opinion that there is a rather basic cultural uniformity in discriminatory reactivity to verbal statements, probably because verbal statements both define and reflect the fundamentals of a relationship between two people.

multidimensionality in ratings of their Depth of Interpretation data, one of their studies, nevertheless, did yield three-dimensionality. The first dimension was clearly one of depth; the second was called Ambiguity; while the third was not identified. The authors later concluded that the second dimension was specific to the raters and materials employed. While the adjectival scale deep-shallow used in the present study is in no real sense equivalent to depth as defined by Raush et al., it is nonetheless interesting to note that such a scale is clearly more highly loaded on the Professional Evaluation factor than on the Precision/Potency factor (see Table 3, Line 37). Were there any relationship between the deep-shallow scale and the defined concept of depth, then such would support the hypothesis that under at least some conditions the depth dimension is perhaps one of evaluation, rather than one of ambiguity or precision. These similarities are, of course, only peripheral. It is also to be noted that Osburn (1951) found no evidence of multidimensionality among his ratings of Ambiguity.

The results confirm the prior impression that raters do indeed tend to react to statements in the semantic differential rating situation with an attitude which is primarily evaluative, and that only in the second place, as it were, do they concern themselves with the degree of ambiguity, clarity, activity, precision, and focus of a response. In complementary fashion, it is a comforting finding that judgments of the second kind are not, after all, necessarily tinged with evaluative considerations. Presumably in some experimental rating situations attributes of evaluation and precision may be confounded and thus heighten error, if the subject is forced arbitrarily to rate along an inadequately defined scale. But happily, the two influences in the present study appear to be quite independent.

Finally, it should fairly be said that Osgood and his colleagues have clearly documented the consistent, primary emergence of an evaluative factor in widely differing contexts—even in the ratings of sonar signals. It is thus in one respect not at all surprising that a similar finding was made in the study reported here, and that a fusion of his activity

and potency factors emerged second. It remains to be seen whether the 10 *therapist responses themselves* largely hold up to a one-, two-, or three-dimensional hypothesis, when rated by three independent groups of subjects upon each of the three dimensions uncovered. Such a study is planned.

SUMMARY

Thirty-five Board-certified psychiatrists rated 10 bona fide therapist verbal responses against 40, seven-point, bipolar adjectival scales, chosen to correspond with the hypothetical variables of Ambiguity, Lead, Inference, Stressfulness, Evaluation, Potency, and Activity. The matrix of intercorrelations among the 40 scales was analyzed by Thurstone's complete centroid method, and the centroid was rotated via the quartimax method, maintaining orthogonality. The first factor, accounting for 33% of the total variance, was one of Professional Evaluation; the second, accounting for 18% of the total variance, was one of Precision/Potency. The third factor discussed (accounting for only 4%) was one of Subjectivity/Objectivity. The results thus indicate that while ratings were made primarily on an evaluation dimension and secondarily on a dimension of precision and potency (ambiguity), the two dimensions are independent. Results are discussed from the standpoints of their limitations and generalizability, the need for further experimentation, and the findings of other investigators.

REFERENCES

- BORDIN, E. S., CUTLER, R. L., DITTMANN, A. T., HARWAY, N. I., RAUSH, H. L., & RIGLER, D. Measurement problems in process research in psychotherapy. *J. consult. Psychol.*, 1954, 18, 79-82.
- COOMBS, C. H. Mathematical models in psychological scaling. *J. Amer. Statist. Ass.*, 1951, 46, 480-489.
- DEUTSCH, F., & MURPHY, W. F. *The clinical interview*. Vol. 2. New York: International University Press, 1955.
- FINESINGER, J. E. Psychiatric interviewing: Principles and procedure in insight therapy. *Amer. J. Psychiat.*, 1948, 105, 187-195.
- FISHER, S. Plausibility and depth of interpretation. *J. consult. Psychol.*, 1956, 20, 249-256.
- FRUCHTER, B. *Introduction to factor analysis*. New York: Van Nostrand, 1954.
- HARWAY, N. I., DITTMANN, A. T., RAUSH, H. L., BORDIN, E. S., & RIGLER, D. The measurement of

- depth of interpretation. *J. consult. Psychol.*, 1955, 19, 247-253.
- HOWE, E. S., & POPE, B. An empirical scale of therapist verbal activity in the initial interview. *J. consult. Psychol.*, 1961, in press.
- NEUHANS, J. O., & WRIGLEY, C. F. The quartimax method: An analytic approach to orthogonal simple structure. *Brit. J. statist. Psychol.*, 1954, 7, 81-91.
- OSBURN, H. G. An investigation of the ambiguity dimension of counselor behavior. Unpublished doctoral dissertation, University of Michigan, 1951.
- OSGOOD, C. E., SUCI, G. J., & TANNENBAUM, P. H. *The measurement of meaning*. Urbana: Univer. Illinois Press, 1957.
- POPE, B., HOWE, E. S., & FINESINGER, J. E. Quantitative studies of therapist verbal activity. Paper presented at American Psychological Association, Cincinnati, September 1959.
- RAUSH, H. D., SPERBER, Z., RIGLER, D., WILLIAMS, JOAN V., HARWAY, N. I., BORDEN, E. S., DITTMANN, A. T., & HAYS, W. A dimensional analysis of depth of interpretation. *J. consult. Psychol.*, 1956, 20, 43-48.
- SNYDER, W. U. (Ed.) *Group report of a program of research in psychotherapy*. University Park: Pennsylvania State Univer., 1953.
- SPEISMAN, J. C. Depth of interpretation and verbal resistance in psychotherapy. *J. consult. Psychol.*, 1959, 23, 93-99.
- THURSTONE, L. L. *Multiple factor analysis*. Chicago: Univer. Chicago Press, 1947.

(Received June 13, 1960)

NEED VALUE AND EXPECTANCY INTERRELATIONS AS ASSESSED FROM MOTIVATIONAL PATTERNS OF PARENTS AND THEIR CHILDREN¹

FORREST B. TYLER, BONNIE B. TYLER, AND JANET E. RAFFERTY

Southern Illinois University

The nature of the relation between need value (goal value) and expectancy has been approached in two somewhat different ways. In more controlled laboratory studies (Crandall, Solomon, & Kellaway, 1955; Edwards, 1955; Feather, 1959; Irwin, 1953; Marks, 1951; Worell, 1956), the subject's preference for known alternatives or the subject's willingness to bet on an outcome with a controlled actuarial probability of occurrence has been used as a measure from which expectancy, need value, and the relation between them have been inferred. In more broadly conceived research such as that of Atkinson and his colleagues (1958), story themes have been used to study strength of a single motive such as need for achievement. Results from both of these types of studies have demonstrated the importance of the two concepts of need value and expectancy in dealing with a variety of problems ranging from predictions of animal behavior in choice situations to analysis of thematic test responses in personality research (Feather, 1959).

Among theoretical approaches to the study of need value and expectancy there have been differences as to the relation between these two variables. Rotter (1954) considers need value and expectancy to be independent, as does Edwards (1955) in his *SEU* model. In contrast, Atkinson (1958) and others (Feather, 1959) consider them interrelated. Empirical evidence indicates that they are statistically interrelated under some conditions, but that the relation found varies for different prob-

ability levels (Crandall et al., 1955) and for different goal value levels (Worell, 1956). Further, any evident relation decreases when a "premium" (i.e., a high goal value) is placed on accuracy of estimates of occurrence (Crandall et al., 1955; Worell, 1956). Also, the relation may be less direct in adults than in children (Crandall et al., 1955; Irwin, 1953; Marks, 1951). Finally, the nature of the relation may be different for achievement and nonachievement situations (Atkinson, 1958; Feather, 1959; Marks, 1951; Worell, 1956).

The objective of the present paper is to present findings pertinent to three questions concerning need value-expectancy relations that are of theoretical as well as practical importance. These questions concern: the independence of need value and expectancy, the relation between need value and expectancy in parents in contrast to that relation in children, and the relation between need value and expectancy for recognition-status motives in contrast to other motives.

METHODOLOGICAL PROCEDURES AND RATIONALE

Subjects

The subjects were three samples of children, aged 2-6 to 5-0, enrolled in a cooperative preschool located on a college campus, and the parents of those children. The three samples included, respectively, 18, 16, and 11 children; total *Ns* were 45 children (20 boys, 25 girls), 45 fathers, and 45 mothers. Forty-three of the 45 sets of parents were college student or faculty families; the remaining 2 families were from the local community.

Collection of Data

Parent data were collected in a 100-question free-response interview structured to elicit information

¹ This investigation was supported by a research grant, M-1137, from The National Institute of Mental Health, United States Public Health Service and by research funds from Southern Illinois University.

concerning parental child-rearing motivations. Each interview required approximately 2 hours to complete and was tape recorded. Interviews were conducted by two trained interviewers, each of whom interviewed half the fathers and half the mothers, but never both members of the same family. Child data were obtained by two trained observers who made narrative records of each child's behavior for 5-minute periods during regular preschool activities. Each child was observed approximately 300 minutes during the first 3 months of school in the fall, with observations for each child distributed equally over the 3-month period. The children were observed in a systematic predetermined manner designed to prevent selective biases. All parent interviews and child observations were typed and data analyses were made from typescripts.

Rating Procedures

The present study has involved categorizing and quantifying parent interview responses and child preschool behaviors on the same motivational variables, while at the same time avoiding contamination in the assessment of parental and child protocols. Accordingly, two research teams operated independently—one working with the child data and one with the parent data—to develop explicit operational definitions for all concepts, and to construct scoring-by-example manuals for the parent data and for the child data using these definitions.

The task of constructing parallel operational definitions of need value and expectancy for parents and for children's behavior was accomplished as follows. All records were analyzed by referents, or behavioral units. Each unit contained information concerning: (a) the stimulus context at the moment (interview question for parents; nursery setting or specific behavioral stimuli for children); (b) the goal-direction of the response (for the parent this goal-direction was inferred from content of verbal statements; for the child it was inferred from the nature of his behavior, e.g., friendly behaviors are considered to be love and affection goal-directed); and (c) other characteristics of the response (e.g., for the parent, persistence, statement of anticipation or dread as accompanying a behavior, etc.; for the child, persistence, constructive-nonconstructive or defensive nature, accompanying verbalization, etc.).

The motivational categories used for classifying these data with regard to goal-directional characteristics are derived from Rotter (1954). As utilized in the present study, they have been given the following operational definitions:

Recognition-Status (R-S): Parents—concern with child's achievements, teaching skills to child, being known as a good parent, doing what a parent should. Children—calling attention to achievement, behaviors, attributes, or possessions; conformity to or imitation of achievement or other socially approved behaviors; attention-getting activities.

Love and Affection (L&A): Parents—concern with play or companionship with child, interest in child's

happiness. Children—cooperative play, friendly behaviors, sharing, helping, sympathetic behavior, seeking affection.

Dominance (Dom): Parents—concern with controlling child, teaching obedience, molding child. Children—commanding others, making demands, aggression, controlling others' activities.

Protection-Dependency (P-D): Parents—(not relevant to child-rearing motivations of parents with preschool children). Children—seeking help, information, permission, comfort or consolation, and intervention of others to prevent frustration.

Independence (Ind): Parents—(not relevant, since independence satisfactions by definition are self-mediated). Children—individualized activity, self care.

Rationale

The following general rationale was developed as a basis for differentiating expectancy and need value referents for each motivational category.

Need value. In general, there is agreement among psychologists that there is a direct relationship between the value of a goal or reinforcement and choice preference for that goal. Accordingly, as a basis for inferring need value (NV) ratings on the seven-point scale used, the following operations were set up.

1. Strength of NV rating is determined by:

a. Variations in stimulus cues as follows: For both parents and children the fewer the stimulus cues, or the less the "stimulus pull," the higher the need value rating given. So if there are very few stimulus cues present—as when an interview question is directed toward eliciting a response in one need area and the response given is directed toward another need—the response is given a relatively high NV rating.

b. Variations in response characteristics as follows: For children, persistence of response leads to a higher rating. For parents, statements concerning persistence, or statements concerning instigation of the indicated kind of activity in situations with few cues present lead to a higher rating.

2. Need category to be scored for NV is indicated as follows:

a. Goal direction² of response, either interview statement or observed behavior, is scored.

b. Goal direction indicated by the situational structure is scored if the situation is "maximized" for response in that direction. This criterion leads to scoring for nonoccurrence of a response when a maximally structured situation is rejected completely by the person responding; this rejection is considered evidence of very low NV in that area. The response direction is also scored as indicated earlier and given a high NV rating if it occurred when there were very few stimulus cues present to elicit it.

This procedure of scoring NV in an inverse relation to the strength of eliciting stimuli is roughly comparable to procedures followed in projective test-

² For a more detailed account of determination of goal-direction of responses, see Tyler (1960).

ing in which attempts have been made systematically to reduce clarity of stimulus cues as a basis for getting at important personality dimensions, and in which unusual or atypical responses are considered most significant (Lindzey, 1952). Scoring nonoccurrence of responses for low NV under specified conditions is less generally accepted, but seemed justifiable and consistent with the broader rationale on which NV strength was determined. The behavioral variable used, persistence, is traditionally considered an indication of the strength of motivation to achieve a goal, and consequently of the value of the goal to the person behaving.

Expectancy. There is considerably less agreement among psychologists as to the relationship between expectancy (Ex) or subjective probability and choice preference for a goal. One point of view is that there is a direct relation between these two (Atkinson, 1958). A somewhat conflicting view is that the effect of low Ex of goal attainment is to lead to the occurrence of "defensive" behaviors with regard to important goals (Eriksen, 1950; Rotter, 1954). Much attention has been given in personality research to study of defenses, even to the extent at times of assuming that nonoccurrence of pertinent behavior must be defensiveness (Zuk, 1956). This extreme position is not taken here. The position is taken that defensive behaviors are not categorically different from other behaviors; rather they indicate the low end of a continuum of constructiveness of goal directed activity. That is, the constructiveness of behaviors is considered to reflect directly the subject's level of subjective probability. Consequently the following measures of expectancy were set up.

1. Need category to be scored for Ex is indicated by goal direction of response.

2. Strength of Ex rating is indicated by:

- a. Variations in response characteristics as follows: For children, ratings range from lowest for withdrawal behaviors, clearly defensive behaviors (e.g., disruptive or nonsocially approved behaviors), and tentative abortive behavior attempts; up the scale to highest ratings for direct, socially approved interactions which include indicators of high confidence (e.g., direct smiling approaches to new children, etc.). For parents, ratings range from lowest for statements of dismay, loss of self-control, and failure and frustration over attempts to attain goals; up the scale to highest ratings for statements of anticipation of interacting with children, joy at interactions together, etc.

- b. Variations in stimulus cues also affect Ex as follows: For children, since data used were records of direct behavior in relatively free situations (i.e., no "forced choices" of a controlled sort existed as they might in experimental situations), it was thought that the levels of expectancy which were quite high might not be differentiable except in situations where a few stimulus cues were present. Consequently it was decided to give the highest ratings only to high confidence responses which occurred in situations where there were a minimum of relevant stimulus cues. Although this method of scoring child Ex makes pos-

sible a slight overlap between child NV and Ex scores, it was deemed essential for the reasons indicated. For parents it was possible to structure questions to elicit expectancy statements independent of need value, so no comparable variations in scoring were required for parental expectancy measures.

Correlations reported are Pearson product-moment r 's (McNemar, 1955) based on mean scores computed from all the ratings given to an individual. Differences between correlations are assessed using Fisher's z' transformation (McNemar, 1955).

FINDINGS

This study has been concerned with charting relationships among motivational characteristics in parents and in children. For that reason findings obtained are meaningful for testing hypotheses primarily in a construct validity fashion (Cronbach & Meehl, 1955). Correlations reported are examined to determine whether they are consistent with relationships hypothesized to exist from specified theoretical points of view. Their support for any such aspects of a "nomological net" constitutes construct validation.

Interrater reliabilities for the parent and child scoring manuals are reported in detail elsewhere (Rafferty, Tyler, & Tyler, 1960; Tyler, Tyler, & Rafferty, 1959). Those figures can be summarized by noting that median interrater reliabilities for these manuals range from .68 to .84 for samples of 18 and 16 subjects.

Correlations reported will be for the total 45 subjects from all three samples studied. The indicated analyses have been run for the separate samples, but the variability in findings between samples is within the limits expected by chance. Hence, they are not reported here.

The need categories used in this study are Recognition-Status (R-S), Love & Affection (L&A), Dominance (Dom), Protection-Dependency (P-D), and Independence (Ind). For each category, both Need Value (NV) and Expectancy (Ex) scores have been obtained.

It was necessary to assess the independence of the need categories to determine whether separate NV-Ex interrelationship analyses were justifiable. The pertinent interneed correlations among NV measures can be summarized briefly. For fathers, mothers, and

children the R-S and L&A NVs are clearly independent, and there is a moderate positive relationship ($r = .40 +$) between R-S and Dom NVs in all three sets of measures. A somewhat smaller but statistically significant inverse relationship ($r = -.30 +$) exists between L&A and Dom NV measures for parents. For girls, this L&A-Dom NV relationship also tends to be negative, though for boys it is positive. However, neither of these correlations approaches statistical significance. Child interneed comparisons do yield marginally significant (.05 level) r 's between R-S and P-D ($r = .34$) and between L&A and Ind ($r = -.31$). Nevertheless, 7 of the 10 interneed correlations for all children are nonsignificant; 8 of the 10 for girls alone are nonsignificant; and all 10 are nonsignificant for boys. In general it seems appropriate to conclude that these correlations indicate sufficient overall independence among these NV measures to warrant consideration of them as functionally distinct motivational categories.

When interneed correlations among expectancy measures are examined more indication of a consistent pattern of positive interrelationships is found than was the case among need value measures. However, this pattern seems to be primarily a function of a generalized level of Ex among the R-S, L&A, and Dom needs which holds for mothers and girls. This generalized relationship is moderate ($r = .40 +$) for mothers, and moderate to high (r 's range from .55 to .75) for girls. In contrast, for fathers the only interneed Ex relationship is a moderate one ($r = .45$, significant at .01 level) between L&A and Dom,

while for sons these three Ex measures are not significantly interrelated. The nature of the sex differential for this generalized Ex pattern is indicated even more clearly by the fact that boy-girl differences are significant (two at the .10 level, one L&A-Dom at the .05) for these three comparisons, and father-girl differences are significant on two of them (R-S-L&A r 's at .05 level, L&A-Dom r 's at .01 level). The differences between mothers interneed Ex correlations and those of fathers and boys also tend to indicate a more generalized Ex for mothers, but none of the correlational differences is significant between mothers and sons, and only one (L&A-Dom, D significant at .05) is significant between mothers and fathers. Even though this somewhat generalized Ex for females is noted, it should also be pointed out that 6 of the 10 interneed expectancy r 's for girls are not significantly interrelated; nor are the 3 Ex measures for mothers so interrelated as to warrant considering them as 3 measures of the same variable. For fathers, only 1 of the 3 comparisons yields even a moderate r , and for boys only 1 of 10 (Dom-Ind $r = .46$, significant at .05 level) achieves significance. Thus there seems to be sufficient independence among these expectancy variables to justify consideration of them as operationally independent.

Results of analyses of NV-Ex interrelationships are reported in Table 1. Findings pertinent to answering the first general theoretical question of the independence of NV and Ex can be summarized as follows:

1. Fathers' NV and Ex scores show a slight

TABLE 1

INTRANEED CORRELATIONS BETWEEN NEED VALUE AND EXPECTANCY MEASURES

Need Category	Parents		Children		
	Fathers ($N = 45$)	Mothers ($N = 45$)	Boys ($N = 20$)	Girls ($N = 25$)	Total ($N = 45$)
Recognition-Status	.19	.00	.23	.60	.40**
Love and Affection	.27	.41**	.22	-.13	.00
Dominance	.21	.07	.71**	.79**	.76**
Protection-Dependency			.66**	.62**	.64**
Independence			.82**	.62**	.72**

Note.—Level of significance: $N = 45 \quad 25 \quad 20$
 $* r_{.05} = .29 \quad .40 \quad .44$
 $** r_{.01} = .40 \quad .51 \quad .56$

TABLE 2
DIFFERENCES BETWEEN NEED VALUE-EXPECTANCY INTERCORRELATIONS
AMONG SUBJECT GROUPS

Need Category	Fa-Mo	Fa-Child	Fa Boy	Fa-Girl	Mo-Child	Mo Boy	Mo-Girl	Boy-Girl
Recognition-Status NV-Expectancy	.192	.232	.042	.501	.424	.234	.693**	.459
Love and Affection NV-Expectancy	.159	.277	.053	.408	.436*	.212	.567*	.355
Dominance NV-Ex- pectancy	.144	.782**	.673*	.857**	.926**	.817**	1.001**	.840
Protection-Dependency NV-Expectancy								.068
Independence NV- Expectancy								.432

Note.—Fathers ($N=45$), Mothers ($N=45$), Boys ($N=20$), and Girls ($N=25$).

* Significant at or beyond the .05 level.

** Significant at or beyond the .01 level.

positive interrelation of the order of .20+, but they are not sufficiently interrelated for that fact to be significant statistically.

2. For mothers, any NV-Ex relation is confined to a moderate one ($r = .41$, significant at .01 level) within the L&A need category.

3. For children, there is a high positive intercorrelation in three need categories, Dom, P-D, and Ind. There is also a high NV-Ex r (.60, significant at .01 level) for girls on the R-S variable. However, NV and Ex measures for children on L&A are completely independent. It would seem that the NV-Ex interrelations obtained are relatively specific, and are a function of at least the following factors: (a) age of subject, e.g., parents show a closer NV-Ex relation on L&A than do children; (b) sex of subject, e.g., boys and girls differ in degree of NV-Ex relationship on R-S; and (c) motivational category under consideration, e.g., boys and girls have a quite different NV-Ex relationship on L&A than on Dom.

The second question of whether NV and Ex are less independent in children than in adults can be answered by reference to Tables 1 and 2. For girls, it can be seen that their need value-expectancy scores are significantly more closely intercorrelated for R-S motives (girl-mother difference significant at .01, girl-father at .06) and for Dom motives (both comparisons significant at .01 level) than is the case

for their parents. However, the converse is true for L&A motives since the NV-Ex r for mothers is significantly greater (.05 level) than that for daughters, and the comparable r for fathers is greater than that for daughters, though this latter difference does not reach an acceptable level of statistical significance. For boys, a similar pattern exists though the differences are not as marked. The greater NV-Ex interrelation on R-S and Dom holds for boys in relation to parents, but is statistically significant only for the latter. Also, the greater NV-Ex interrelation for parents on the L&A category holds with respect to boys and parents, but it does not achieve significance.

These comparisons of parental and child NV-Ex relations do not indicate consistently closer NV-Ex relations for children; rather, they seem to indicate that whether parents or children will be found to have a closer NV-Ex relation is a function of the motivational category under consideration.

The third question asked is that of the relation between NV and Ex for R-S motives in contrast to other motives. The hypothesis for which this comparison is pertinent is one advanced by Atkinson (1958) which states that there is an inverse relationship between incentive value and subjective probability for need Achievement which does not maintain

for other needs. Data to test this hypothesis are the relationships presented in Table 1, with the pertinent comparisons being those between rows for each subject group. For fathers there are no differences from need to need in the NV-Ex relation. For mothers, the R-S and Dom NV-Ex relations are not different, though the comparable L&A relation is significantly greater than comparable comparisons on the R-S variable ($D = .436$, significant at .05 level), and the Dom variable ($D = .366$, significant at .10 level). For children the NV-Ex relation on R-S is intermediate in relative as well as absolute magnitude. It is greater ($D = .424$, significant at .06 level) than the comparable relation for L&A and less than that on the other three need variables (P-D $D = .334$, nonsignificant; Ind $D = .484$, significant at .05 level; Dom $D = .572$, significant at .01 level). The findings are not consistent with the hypothesis advanced by Atkinson since for none of the three groups is there clearly a less direct or more inverse relationship between NV and Ex for R-S motives than for all other motives.

DISCUSSION

The three theoretical questions concerning need value-expectancy interrelationships which are the primary focus in this article concern (a) the independence of need value and expectancy, (b) the greater independence of need values and expectancies in adults than in children, and (c) the less direct relationship between need value and expectancy variables for recognition-status motives than for other motivational categories. It should be noted that previous data from which answers to these questions have been inferred have frequently been derived in controlled experimental situations where known expectancies and goal preferences could be systematically manipulated. In contrast, the findings of this study are based on more molar data with mean scores for each subject derived from ratings of a number of referents (answers to interview questions, or observed behaviors). Although it is not appropriate to assess from the present findings the validity of results obtained in specific experimental settings, it is possible to determine from these results the representativeness of those more molecular

situations and the generality of conclusions derived from them.

The question of a general need value-expectancy relation which holds for all needs and all subjects has been raised by Rotter (1954), Atkinson (1958), Edwards (1955), and others (Feather, 1959). Neither a general position of NV-Ex independence (Edwards, 1955; Rotter, 1954) nor a position of interdependence (Atkinson, 1958) can be supported by the findings of this study. Rather, it would seem that need value-expectancy interrelations are specific to the sex and age of the subjects studied, and to the motivational category measured. Although fathers yield a stable nonsignificant NV-Ex relationship for all need comparisons, none of the other subject groups do. Further, the NV-Ex relationship for parents is different from that for children on both L&A and Dom, and this interrelation for girls on R-S is significantly different from that for boys, for fathers, and for mothers on the R-S variable. In general, these parent-child differences in need value-expectancy relationship such as that obtained on the L&A variable support the conclusion that adult motivational patterns are not clearly established during the preschool years. Further, the obtained sex differences, such as that between boys and girls on R-S, provide substantial evidence of sex-linking in patterns of motivational development. The specificity of relations found leads to underscoring the need for caution in generalizing from the results of one study to motivational characteristics of the population at large.

The second question which concerns a closer relationship hypothesized between need values and expectancies in children than in parents is derived from a suggested explanation for discrepancies in results obtained by Crandall et al. (1955) and Irwin (1953) in contrast to those of Marks (1951). Crandall has indicated that the smaller effect of reinforcement values on expectancy statements found in his work and that of Irwin, in comparison to Mark's findings, may be a function of greater susceptibility to reinforcement value effects in younger children, since Marks used 9 to 11 year olds as subjects, and Crandall and Irwin used college students. If such a phenomenon does exist, then preschool children's NV and

Ex measures should be more closely related than is the case for adults (i.e., their parents). The results obtained suggest that this explanation is an oversimplification of the relevant variables. Specifically, on the Dom variable the children show a greater NV-Ex interrelationship than the parents do, on the L&A variable the parents show a greater relationship than the children, and on the R-S variable the boys show a similar NV-Ex relationship to that of their parents, although the girls show a closer relationship.

The third theoretical question concerns the hypothesis advanced by Atkinson (1958) that there is an inverse relationship between incentive value and subjective probability for *n* Ach which does not maintain for other needs. As noted earlier there is not complete overlap between the R-S motive category and the *n* Ach category, but R-S would seem to be subsumed within the latter category. Consequently, findings concerning it should be pertinent. Findings from the present study show that the indicated R-S NV-Ex correlation is not significantly closer to an inverse relationship for R-S than for all other needs for any of the four groups (fathers, mothers, boys, girls) analyzed. In fact, for each group of subjects there is at least one other variable on which the NV-Ex relationship is no different from or less direct than the R-S NV-Ex relationship. Therefore, although these findings support an assumption of specificity of NV-Ex interrelationships to age and sex of subjects, and to motivational category under study, they are not consistent with Atkinson's hypothesis. It may be that this discrepancy is a function of the lack of comparability between the more molar level of measurement in the present study and the more specific and controlled measurement situations in which his work has been conducted. This discrepancy may also suggest that the inverse relationship between subjective probability and incentive value which seems to hold in his experimental situations may not be an important determinant of behavioral choices of the sort on which the findings of the present study are based. However, the obtained intersubject differences on the R-S variable warrant particular consideration, since many investigations of motivation have focussed on achievement-

type situations and tasks. Although the recognition-status motive includes only other-mediated achievement goals, so it is not completely equivalent to *n* Ach, it does fall within the achievement category. To the extent that these concepts overlap, the present findings indicate clearly that for preschool girls importance of achievement goals and subjective probability of attaining those goals are much more closely linked than is the case for preschool boys or for parents. Since this differential does not hold for the mothers' child-rearing motivations, it would seem that girls may acquire this distinction at a later date than is the case for boys. Nevertheless, these findings tend to confirm the results of Veroff (1953), which indicate that there are significant sex differences in response to achievement situations.

SUMMARY

In the present paper relations among the need values (NVs) and expectancies (Ex's) of child-rearing motivations of parents and of motivations of their preschool children are analyzed. Subjects were 45 families with children enrolled in a cooperative preschool in a university housing project. Parental motives for recognition-status (R-S), love and affection (L&A), and dominance (Dom) were assessed from free-response interviews; child motivations for the same three motives, and for protection-dependency (P-D), and independence (Ind) were assessed from narrative observations of preschool activities. Adequate interrater reliabilities for both parent and child score-by-example manuals for rating the indicated protocols are reported.

Patterns of intercorrelation among the motivational variables measured support the conclusion that they are operationally independent. Need value-expectancy correlations obtained are viewed as evidence in a construct-validation sense of the validity of certain hypotheses which have been advanced by others concerning theoretical interrelations between need value and expectancies. On the basis of these findings, there is no support for the hypothesis that there is either a general pattern of independence or a general pattern of interrelation between NVs and Ex's which holds for all subjects and all motivational cate-

gories. Rather, there is support for the position that NV and Ex interrelations are specific to the need category under study, and to age and sex of subjects. The hypothesis that NVs and Ex's are less independent in children than in adults is likewise not supported generally, since this interrelation also seems to be a function of the motivational category under study, with parents differentiating more clearly on one of the three common motivational categories, and children differentiating more clearly on one. The hypothesis that there tends to be an inverse relation between NVs and Ex's for R-S motives which does not maintain for other motives could not be supported. Although there were clear NV-Ex correlation differences from need category to need category, for none of the subject groups was the R-S NV-Ex r clearly less direct than was the case for the other motives.

REFERENCES

- ATKINSON, J. W. (Ed.) *Motives in fantasy, action, and society*. Princeton, N. J.: Van Nostrand, 1958.
- CRANDALL, V. J., SOLOMON, D., & KELLAWAY, R. Expectancy statements and decision times as functions of objective probabilities and reinforcement values. *J. Pers.*, 1955, 24, 192-203.
- CRONBACH, L., & MEEHL, P. E. Construct validity in psychological tests. *Psychol. Bull.*, 1955, 52, 281-302.
- EDWARDS, W. The prediction of decisions among bets. *J. exp. Psychol.*, 1955, 50, 201-214.
- ERIKSEN, C. W. Some implications for TAT interpretation arising from need and perception experiments. *J. Pers.*, 1950, 19, 282-288.
- FEATHER, N. T. Subjective probability and decision under uncertainty. *Psychol. Rev.*, 1959, 66, 150-164.
- IRWIN, F. W. Stated expectations as functions of probability and desirability of outcomes. *J. Pers.*, 1953, 21, 329-335.
- LINDZEY, G. Thematic Apperception Test: Interpretive assumptions and related empirical evidence. *Psychol. Bull.*, 1952, 49, 1-25.
- McNEMAR, Q. *Psychological statistics*. New York: Wiley, 1955.
- MARKS, ROSE W. The effect of probability, desirability, and "privilege" on the stated expectations of children. *J. Pers.*, 1951, 19, 332-351.
- RAFFERTY, JANET E., TYLER, BONNIE B., & TYLER, F. B. Observations of free play situations as a method of personality assessment. *Child Develpm.*, 1960, 31, 691-702.
- ROTTER, J. B. *Social learning and clinical psychology*. New York: Prentice-Hall, 1954.
- TYLER, BONNIE B., TYLER, F. B., & RAFFERTY, JANET E. A systematic approach to interviewing as a method of personality assessment. Paper read at American Psychological Association, Cincinnati, September 1959.
- TYLER, F. B. A conceptual model for assessing parent-child motivations. *Child Develpm.*, 1960, 31, 807-815.
- VEROFF, J., WILCOX, S., & ATKINSON, J. W. The achievement motive in high school and college age women. *J. abnorm. soc. Psychol.*, 1953, 48, 108-119.
- WORELL, L. The effect of goal value upon expectancy. *J. abnorm. soc. Psychol.*, 1956, 53, 48-53.
- ZUK, G. H. The influence of social context on impulse and control tendencies in preadolescents. *Genet. psychol. Monogr.*, 1956, 54, 117-166.

(Received June 13, 1960)

CROSS-VALIDATION OF A RORSCHACH CHECKLIST ASSOCIATED WITH SUICIDAL TENDENCIES

IRVING B. WEINER¹

University of Rochester School of Medicine and Dentistry

Although numerous methods for assessing suicidal potential by means of the Rorschach test have appeared in the literature, failure to cross-validate "signs" and "configurations" supposedly indicative of suicidal tendencies has cast doubt on the utility of these methods (Fisher, 1951; Sakheim, 1954). However, recent work by Daston and Sakheim (1960) using Martin's checklist, 17 signs empirically derived from comparisons between suicidal and nonsuicidal psychiatric patients, has yielded promising results. Daston and Sakheim compared the Rorschach protocols of patients who were nonsuicidal with those of patients who had attempted suicide or had actually taken their own lives. In their study there were virtually no differences between completed suicide and suicide attempt groups on Martin's checklist, but both of these groups received significantly more of Martin's signs than the nonsuicidal group. Furthermore, 83% of the successful suicides and 72% of the attempted suicides received six or more of Martin's signs, while only 17% of the controls displayed this many signs.

Since the above results were derived from a population of hospitalized male veterans undifferentiated with respect to psychiatric diagnosis, question may be raised concerning the concurrent validity of Martin's checklist for a population less homogeneous than the veteran group. This question in turn points to the importance of evaluating the influence of such variables as age, sex, hospitalization, and nature of psychopathology on the checklist scores. Additionally, since most Rorschach scores are significantly related to the total number of responses given (Fiske & Baugh-

man, 1953), it is pertinent to examine the relationship between response total and checklist score.

PROCEDURE

Two samples were used in this study. The first, which was intended to provide general information about Martin's signs, consisted of all adult patients in a 6-month period from the psychiatric services of a general hospital for whom scorable Rorschach protocols were available. Patients whose primary diagnosis was organic rather than functional in nature were excluded. The sample contained 28 males and 43 females, had a median age of 29.5 years (range 15-55), and was comprised of 42 hospital and 29 clinic patients. The total group is categorized by age, sex, and patient status in Table 1.

TABLE 1
AGE, SEX, AND PATIENT STATUS OF SUBJECTS

Age	Number Males		Number Females		Total
	Hospital	Clinic	Hospital	Clinic	
15-21	1	4	6	5	16
21-30	4	5	11	4	24
31-40	6	4	7	4	21
41-55	3	1	4	2	10
Total	14	14	28	15	71

The Rorschach records of these 71 patients, which ranged in number of responses from 7 to 67 with a median of 20.33, were scored for Martin's signs, which are the following: No. $D < 6$ or > 20 ; $D\%$ < 60 or > 79 ; No. $CF > 0$ to < 3 ; Total Color $R < 1$; $Sum C > 1.0$ to < 3.5 ; C or CF appear first on $VIII-X$; C or CF with $Sum Y + Sum T < 1$; No. $FV + VF < 1$; $Sum Y < 1$; $Sum Y + Sum T < 1$; difference between M and $Sum C < 1.5$; No. $H + Hd > 6$; No. Categories < 6 or > 13 ; $VIII-X/R > 29\%$; No. $P < 3$ or > 6 ; $P < 3$ with $F + \% > 60$; and time first $R < 27$ seconds.

Subsequently the subjects' hospital and clinic records were utilized to assign the subjects to one of three diagnostic categories: *neurosis*, which included cases of conversion hysteria, obsessive-compulsive

¹ Appreciation is expressed to Norman I. Harway for suggestions concerning the manuscript.

neurosis, anxiety reaction, and neurotic depressive reaction; *character disorder*, which applied to instances of personality trait and personality pattern disturbance; and *psychosis*, which included schizophrenic, psychotic depressive, and involutional psychotic reactions. The diagnostic criterion was the label assigned to the patient during that psychiatric contact in which the Rorschach had been given. For hospital patients, these labels were the discharge diagnoses; for clinic patients, the recorded consensus of an intake committee was used. The hospital group contained 13 neurotics, 14 character disorders, and 15 psychotics; for the clinic group, the respective totals were 13, 10, and 6.

Further examination of the records revealed that eight of the subjects had made documented, serious suicide attempts which had precipitated their admission to the hospital. A search of case files for the year previous to the 6-month period of the study yielded an additional sample of 16 patients who had received psychological testing during a hospital admission subsequent to an attempt at suicide. Since this second group had by and large been referred for testing from the same wards by the same personnel and had been evaluated by the same psychologists as the original eight suicide attempts, it was felt appropriate to combine these two groups for purposes of comparison with the 63 nonsuicidal patients originally studied. The suicidal group was composed of 9 males and 15 females, had a median age of 30 years (range 16-55), and had given between 7 and 57 Rorschach responses with a median response total of 18. Seven of these patients had been diagnosed neurotic, nine as character disorders, and eight as psychotic. It may be noted that this rather even nosological distribution of suicidal patients is consistent with the findings of Pokorny (1960). The observed similarity between completed suicide and suicide attempt groups (Daston & Sakheim, 1960) appears to justify non-inclusion of actual suicides in this study. The significance of suicide attempts as a precursor of actual suicide, as discussed by Shneidman and Farberow (1957, pp. 3-10) and Robins, Gasner, Kayes, Wilkinson, and Murphy (1959), further supports this decision.

RESULTS

The original 71 subjects received between 4 and 12 of Martin's signs, with a median of 7.29 signs. As the distribution of numbers of signs did not approach normality (Rorschach scores in general are not normally distributed [Fiske & Baughman, 1953]), and as the nature of the sampling did not provide equal numbers of subjects for various subgroupings, the data were treated with nonparametric techniques. The distribution of number of Martin's signs was divided into five approximately equal parts for comparison with the variables of age, sex, patient status, and num-

TABLE 2
ASSOCIATION BETWEEN MARTIN'S SUICIDE SIGNS,
SUICIDE CLASSIFICATION, AND DIAGNOSTIC
CATEGORY

Suicide Classification	Diagnostic Category		
	Neurosis	Character Disorder	Psychosis
Suicide attempt:			
Number above Mdn ^a	6	7	6
Number below Mdn ^b	1	2	2
Nonsuicidal:			
Number above Mdn ^a	5	10	11
Number below Mdn ^b	20	9	8

^a Eight or more signs.

^b Fewer than eight signs.

ber of Rorschach responses by means of $5 \times n$ contingency tables. The distributions of age and number of responses were each divided into four equal parts; for the dimension of patient status, hospital (inpatient) and clinic (outpatient) groups were contrasted. None of the obtained χ^2 values for the association between number of Martin's signs and these four variables approached significance.

In Table 2 are listed the frequencies of suicidal and nonsuicidal patients with different diagnoses who received more or less than the total group median number of Martin's signs. Analysis of these data by Wilson's (1956) method for nonparametric analysis of variance revealed the following: (a) number of signs and suicide classification were significantly associated beyond the .005 level of confidence; (b) the association between number of signs and diagnostic category was significant beyond the .05 level; and (c) there was no significant interaction between suicidal classification and diagnostic category.

The median numbers of signs received by the two groups were 8.83 for the suicide attempts and 7.04 for the controls. A median test indicated that the discrepancy between these values was significant at the .01 level of confidence. No clear cutting score separating suicidal from nonsuicidal patients emerged from the data; the most efficient cutting score, an incidence of eight or more signs, correctly classified 79% of the suicide attempts and 60% of the controls.

The 17 signs were examined individually for their capacity to differentiate between the suicidal and nonsuicidal groups. A series of χ^2

tests revealed that two signs (C or CF appear first on VIII-X and $P < 3$ with $F + \% > 60$) had been received by significantly ($p < .05$) more of the suicidal than the nonsuicidal group; trends ($p < .10$) in this direction occurred with two signs (No. $CF > 0$ to < 3 and C or CF with $Sum Y + Sum T < 1$); one sign (VIII-X/R $> 29\%$) significantly differentiated the two groups in the direction opposite from expectation; and the other signs did not significantly discriminate suicidal from nonsuicidal patients.

DISCUSSION

The data indicate that, within the ranges sampled by this study, scores on Martin's checklist may be interpreted similarly for men and women patients of different ages, whether they are hospitalized or applying for outpatient psychiatric care, and regardless of how many responses they give to the Rorschach. Number of Martin's signs is significantly and positively associated both with likelihood of having made a suicide attempt and with severity of psychopathology, if the labels neurosis, character disorder, and psychosis may be taken as a continuum of increasing emotional disturbance. The significance of both main effects, in the absence of any significant interaction, indicates that although the potential of the checklist to assess severity of pathology may be a topic for further investigation, the capacity of the measure to discriminate suicidal from nonsuicidal patients operates independently of diagnostic category.

The operation of the individual signs merits consideration, although the positive findings do not add much to current conceptions of what kinds of people attempt suicide. The best discriminator between the suicide attempts and the controls, $P < 3$ with $F + \% > 60$, may be interpreted as a rejection of conventional behavior patterns in the presence of adequate reality testing. However, this sign, though of predictive value, is too rare even in suicidal patients to have general descriptive value. The other significant discriminator and the two signs which approach significance (see Results) may reflect difficulty in dealing comfortably with affective stimulation and controlling tendencies toward

impulsive behavior. But such speculations do not provide hypotheses about suicidal individuals which have not previously been advanced without recourse to projective test data, and the value of the checklist would seem to lie more in practical application than in contribution to theory. Concerning the several signs which contributed little or nothing to the differentiation between suicidal and nonsuicidal groups, the moral may be iterated that empirically derived items require careful and repeated cross-validation before they achieve status as efficient and reliable predictors.

With regard to future use of Martin's checklist, a final point should be mentioned. Rosen (1954) and Meehl and Rosen (1955) have questioned the efficiency of psychometric instruments for the prediction of rare events such as suicide and argue that the inevitable number of false positives in such endeavors renders even highly valid discriminators impracticable. However, Cureton (1957) has discussed a method of allowing for population base rates in the establishment of cutting scores to predict a dichotomous criterion from a continuous predictor. The necessity is therefore for collection of sufficient normative data to establish the distribution of the predictor, i.e., Martin's checklist, in order that through consideration of suicide rates, efficient prediction of suicidal risks may be implemented. The data of the current study would suggest that the checklist has adequate concurrent validity to justify such future research.

SUMMARY

Rorschach protocols of 71 patients who had been tested during a 6-month period were scored for Martin's suicide signs. The number of signs earned was found to be relatively independent of the age, sex, hospital status, and Rorschach response total of the subjects. Subsequently, the records of 63 of these patients who were nonsuicidal were compared with those of 24 patients who had made suicide attempts. The suicidal group received significantly more of Martin's signs than the controls. A significant positive association was also found between the number of signs received and severity of psychiatric illness. The

absence of any significant interaction between suicidal classification and diagnostic category indicated that the relationship between suicidal disposition and number of signs was similar in the different diagnostic groups, however. The results did not delineate any efficient cutting scores for predictive use. Nevertheless, it is felt that the demonstrated concurrent validity of Martin's checklist justifies further research to develop reliable norms which, when considered in the light of base rate incidence, might facilitate efficient prediction of suicidal tendencies.

REFERENCES

- CURETON, E. E. Recipe for a cookbook. *Psychol. Bull.*, 1957, **54**, 494-497.
- DASTON, P. G., & SAKHEIM, G. A. Prediction of successful suicide from the Rorschach test using a sign approach. *J. proj. Tech.*, 1960, **24**, 355-361.
- FISHER, S. The value of the Rorschach for detecting suicidal trends. *J. proj. Tech.*, 1951, **15**, 250-254.
- FISKE, D. W., & BAUGHMAN, E. E. Relationships between Rorschach scoring categories and the total number of responses. *J. abnorm. soc. Psychol.*, 1953, **48**, 25-32.
- MEEHL, P. E., & ROSEN, A. Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychol. Bull.*, 1955, **52**, 194-216.
- POKORNY, A. D. Characteristics of forty-four patients who subsequently committed suicide. *AMA Arch. gen. Psychiat.*, 1960, **2**, 314-323.
- ROBINS, E., GASNER, S., KAYES, J., WILKINSON, R. H., & MURPHY, G. E. The communication of suicidal intent: A study of 134 consecutive cases of successful (completed) suicide. *Amer. J. Psychiat.*, 1959, **115**, 724-733.
- ROSEN, A. Detection of suicidal patients: An example of some limitations in the prediction of infrequent events. *J. consult. Psychol.*, 1954, **18**, 397-403.
- SAKHEIM, G. A. Suicidal responses on the Rorschach test: A validation study. *Dissertation Abstr.*, 1954, **14**, 1253-1254.
- SHNEIDMAN, E. S., & FARBEROW, N. L. *Clues to suicide*. New York: McGraw-Hill, 1957.
- WILSON, K. V. A distribution-free test of analysis of variance hypotheses. *Psychol. Bull.*, 1956, **53**, 96-101.

(Received June 14, 1960)

CONCURRENT AND CONSTRUCT VALIDITY OF DIRECT AND INDIRECT MEASURES OF DEPENDENCY

MARVIN ZUCKERMAN¹

Institute of Psychiatric Research, Indianapolis, Indiana

EUGENE E. LEVITT, AND BERNARD LUBIN²

Indiana University Medical Center

Research on acquiescence response set (Zuckerman, Norton, & Sprague, 1958) and suggestibility (Zuckerman & Grosz, 1958) led to an interest in the possible relatedness of these variables to the trait called "dependency." In assessing this trait, peer nominations, objective and projective tests had been used. However, anyone working in the field of personality plunges into chaos when he faces the problem of measurement. One finds that widely differing techniques claim to measure the same hypothetical variable, but in fact do not correlate with each other. The answer to this lack of construct validity given by some assessment theorists (Leary, 1957) is that the tests are measuring "different levels of personality." The concept of levels is related to the idea of a continuum of consciousness-unconsciousness. The more direct tests (those which ask the testee to describe his own feelings and reactions) are presumably at the upper levels, while the less direct tests (those which are disguised as creative or perceptual tasks) tap the lower, "unconscious," levels. This distinction does not resolve the psychometric dilemma, for one must still demonstrate that a test at any level is measuring something more general than itself, even if this is indicated only by a correlation with other measures at adjacent levels. Furthermore, at some point, a test should relate to

behavior as assessed by something other than another test. This may be observations by psychologists in life situations, measurements in controlled miniature situations, or descriptions of the subject made by persons who have had a chance to observe him over some period of time.

The study reported here represents an attempt to bring some order to measurement of one construct: dependency. An attempt was made to: conceptualize the dimensions of the construct; select a battery of tests covering the range of the direct-indirect continuum and representing some of the current methodologies of test development; score these tests within the same conceptual framework; assess their validity in two ways: (a) concurrent validity, by comparing all tests against an external criterion, a rating by peers; (b) construct validity, by factor analyzing the correlations between all tests and peer ratings. It was assumed that the largest factor to emerge from a factor analysis of all variables would represent a broad dependency factor, and that loadings of the tests on this factor could be taken as an indication of their construct validity. Of course, secondary factors also might have some bearing on construct validity. In fact, it was hoped that the factor analysis might help conceptualize the dimensions of the broad construct, dependency, beyond the initial working hypotheses.

The peer rating was selected as the single criterion to use for concurrent validity because it was closest to behavioral description in a general sense. It should be noted that the

¹ Now at Department of Psychology, Brooklyn College, Brooklyn 10, New York.

² The authors are indebted to Beatrice Barrett, Robert A. Wagoner, Harry Brittain, and George Petoe for their assistance in judging and scoring test records.

measurement of personality from this source depends on the effect that the subject's behavior has on her peers, and does not necessarily have a relevance to the subject's own image of herself or her underlying motives and defenses.

METHOD

The concept of dependency used in this study stems from Horney's (1945) description of the "compliant" or "moving-toward-people" personality. Since some of the tests we considered using were scorable within the Murray (1938) need system, it was convenient to translate Horney into Murray. Horney delineated three traits of the compliant personality: a marked need for affection and approval from others (Succorance), a tendency to subordinate himself to others and to inhibit assertiveness and criticality (Deference and Abasement), a tendency toward self-blame and guilt (Abasement). Murray needs at the opposite poles are: Autonomy and Dominance.

Subjects

The initial subjects were 78 sophomore student nurses living together in a dormitory in the Indiana University Medical Center. They were acquainted with one another for about 5 months previous to the time when tests were given. Six of this initial group were later excluded from the study because their tests were incomplete or because their Consistency score on the EPPS was less than 9. The remaining group of 72 subjects were almost all between 19 and 20 years of age. Their average ACE score was 58.6 (national norms percentile), indicating slightly higher than average intelligence, relative to college students. The group's mean scores on the Edwards Personal Preference Schedule variables were compared with Edwards' (1954) means for 749 female college students. Critical ratios higher than 2.00 were obtained for differences on 6 of the 15 scales. The student nurse group scored significantly higher on Abasement, Intracception, Nurturance, and Endurance, and significantly lower on Dominance and Change. The differences on Dominance, Abasement, and Nurturance scores replicate differences found between a previous student nurse group from this school and Edwards' college females (Zuckerman, 1958). Additional comparisons were done on the Navran Dependency scale (1954) and Gough Dominance scale (Gough, McClosky, & Meehl, 1951). The student nurse group scored significantly higher ($p < .001$) on the Navran Dependency scale than a group of 200 normals described by Navran (1954). The average score of the nurse group on the Gough Dominance scale fell between the average scores reported (Gough et al., 1951) for high and low Dominance groups, but fell much closer to the mean for the low Dominance groups. In sum, it would appear that this group of subjects differs from the usual female college student group in various personality

traits. The possible bearing of these differences on our results will be discussed at the end of this paper.

Peer Ratings (PR)

The items from four dimensions of Leary's Interpersonal Check List (Leary, 1957) were adapted to form three eight-point bipolar scales providing description at each point and fitting the hypothesized dimensions of the construct. These scales are given below:

I. Pride-Shame

1. Expects everyone to admire her
2. Always giving advice, acts important, tries to be too successful
3. Makes a good impression, often admired, respected by others
4. Well thought of
5. Able to criticize herself
6. Apologetic, easily embarrassed, lacks self-confidence
7. Self-punishing, shy, timid
8. Always ashamed of herself

II. Dominant-Submissive

1. Dictatorial
2. Bossy, dominating, manages others
3. Forceful, good leader; likes responsibility
4. Able to give orders
5. Can be obedient
6. Usually gives in, easily led, modest
7. Passive and unaggressive, meek, obeys too willingly
8. Spineless

III. Independent-Dependent

1. Egotistical, conceited, cold, unfeeling
2. Boastful, proud, self-satisfied, snobbish, thinks only of herself, shrewd, calculating, selfish
3. Independent, self-confident, self-reliant, can be indifferent to others, business-like
4. Self-respecting, able to take care of herself
5. Grateful, appreciative
6. Often helped by others, admires and imitates others, very respectful, very anxious to be approved of, accepts advice readily, trusting, and eager to please
7. Dependent, likes to be taken care of, easily fooled, wants to be led
8. Clinging vine, will believe anyone

The subjects were asked to rate every nurse in the group whom they felt they knew or had observed enough to rate. They rated a peer by circling one of the numbers from 1 to 8 on each of the three scales. A combination score was obtained by summing each subject's ratings for each peer that she rated.

Tests

Six basic kinds of instruments were used. In order of what was felt to be their directness, we have:

1. Self-Ratings (SR): All subjects made self-ratings on the same scales that they used to make peer ratings.

2. True-False Questionnaires (Q):

a. Gough (1951) Dominance scale, developed from the MMPI using an empirical method, or item selection based on criterion groups

b. Navran (1954) Dependency scale, developed from MMPI using content validity, or item selection by clinical judges

In addition to the scores derived from the separate questionnaires, a combination score suggested by Ullman (1958) was formed by counting items on both tests scored for submissiveness and dependency and eliminating overlapping items. Ullman called this combined scale "Lack of Self-Assertion."

3. Forced-Choice Questionnaire: Edwards Personal Preference Schedule (EPPS) (Edwards, 1954). It is assumed that the elimination of the social desirability factor makes this kind of forced-choice questionnaire less direct than the usual true-false questionnaire, since the subject is less able to choose his responses in conformity to a more obvious stereotype of "adjustment." Furthermore, the pairing of 15 needs probably makes it more difficult for the subject to grasp what is being measured than when all items pertaining to a particular need are presented in one scale. The only scales actually used were those relevant to our construct: Deference, Succorance, Abasement, Autonomy, and Dominance. A combination, or ratio, score was formed by converting the raw scores to Edwards' standard scores and taking the ratio of Deference plus Succorance plus Abasement to the total sum of all five scores.

4. Sentence Completion Test: Rohde's (1957) Sentence Completion Test (SC) was used since the manual describes scoring in the Murray need system. The SC often is classed as a projective test because it is free-response. But it is more direct than other projective tests because it asks the subject to describe his own feelings. The items were scored for the same variables as were scored in the EPPS, and the same ratio score was used as a combination score. The three experimenters scored the test one item at a time over all subjects. An initial discussion of the range of responses on the particular item was followed by independent scoring and then acceptance of scores with a minimum consensus (two out of three agreement on the basic score).

5. TAT: Ten cards from the TAT (Cards 2, 3GF, 4, 6, 7BM, 7GF, 9GF, 10, 12M, 18GF) were presented to the group using an opaque projector. Subjects were allowed 5 minutes per card to write their stories. The three experimenters scored the stories one story at a time over all subjects. Each experimenter gave the story an initial score, and disagreements were resolved by conference technique. The five needs scored were the same as in the EPPS and SC, and a ratio score was formed in the same manner as on these two previous tests.

6. Rorschach: The Rorschach was group administered using an opaque projector to project the cards on a screen. Subjects were allowed 3 minutes to write

down the different things they could see on a card. A scoring system for Dependency content was adapted from that of DeVos (1952). After some practice scoring, it was found necessary to eliminate some of his more ambiguous categories. Two assistants independently scored the content of the 72 records. The correlation between scorings was .83. The frequency scores of the two scorers were averaged for each subject, and this average was divided by the subject's total number of responses to get a Dependency percentage score.

RESULTS AND DISCUSSION

Peer Rating Reliabilities and Intercorrelations

The number of peers rating each subject ranged from 37 to 74 with a median of 57. Reliabilities of the peer ratings were calculated using a formula devised by Horst (1949). The resultant reliability coefficients for PR_I, PR_{II}, PR_{III}, and the Sum PR, were .94, .96, .94, and .96. The correlations between the three PRs were close to the maximum limit set by their reliability: .92, .91, and .92. Apparently, the subjects made no distinctions between the three scales in rating their peers. However, the halo effect cannot be attributed to a general positive-negative reaction because the "negative," or "undesirable," descriptions are at both ends of the scales. A general evaluation might affect degree, but could not affect direction of the ratings, i.e., the rater still had to decide whether she did not like the ratee because the ratee was too dependent or too independent. However, this "across-scales" halo effect precluded using the PR to reveal possible dimensions of the general construct. Therefore, only the Sum PR was used in all comparisons with other measures. This measure was averaged across judges for each subject. The distribution of these average PRs was essentially normal.

Self-Ratings: Intercorrelations

The subjects showed more variation between the three scales in rating themselves, for the correlations between SRs on the three scales were .44, .37, and .47. It was therefore possible to analyze these scales in a combination score and singly.

Analysis of Combination Scores

Table 1 lists the intercorrelations of the sum and ratio scores for each of the basic

techniques. The correlations between each of the tests and the PR can be seen in the first row across. These correlations tend to fall off as a function of the postulated indirectness of the test. The self-ratings yielded the highest correlation, the questionnaires of both types and the SC test were intermediate, and the TAT and Rorschach were lowest. Only the three more direct tests correlated significantly with the criterion, the SC test correlated just below the .05 level, and the TAT and Rorschach correlated close to zero with the criterion. The combined indices did not mask any high relationships between individual scores and the criterion, although in some cases one could have done just as well or slightly better using a single variable. For instance, the Gough Dominance scale correlated more highly ($-.33$) when used alone than when used in combination with the Navran Dependency scale. While the SC ratio did not correlate significantly with the PR, the SC Autonomy score did correlate significantly ($-.33$).

It can also be seen in Table 1 that all of the three most direct techniques were significantly intercorrelated. The SC test correlated significantly with the EPPS, and approached significance in its correlations with the other more direct techniques. The TAT and Rorschach did not correlate significantly with each other or any of the other techniques. Another result seen in this table is the fact that the highest positive correlations are on the diagonal of the matrix and that correlations tend to drop off or become negative in either direction from the diagonal. This means that each test tends to vary most directly with the tests closest to it on the postulated continuum of "directness-indirectness." This finding offers support for the original placement of these tests in this continuum. The SC test seems to have more in common with the direct objective tests than with the indirect free-response tests.

Another analysis of TAT and Rorschach was undertaken in order to check the possibility that a more global approach to the indirect techniques might yield greater concurrent validity. The TAT and Rorschach protocols of the 10 highest and 10 lowest subjects on the Sum PR criterion were selected for this

TABLE 1
INTERCORRELATIONS OF SUM AND
RATIO SCORES

Technique	SR	Q	PPS	SC	TAT	Ror	ACE
PR	.44**	.28*	.24*	.22	.05	.02	-.31**
SR		.37**	.37**	.23	-.22	-.12	-.14
Q			.52**	.18	-.10	-.19	-.25*
EPPS				.39**	.09	-.13	-.30*
SC					.11	-.17	-.27*
TAT						.08	-.10
Ror							.17

* Significant at or below .05 level.

** Significant at or below .01 level.

purpose. These protocols were given to two experienced clinical psychologists along with the rating instructions used by the peers in making their ratings. They were asked to predict which subjects were rated as dependent and which as independent by their peers. They were informed that one-half of the group of 20 fell into each of these categories, and were asked to make their sortings using the same distribution. Quantitative scores involved in the comparisons were dichotomized at the medium and comparisons were made using exact probability tests for 2×2 tables.

Using the TAT, the judges did not demonstrate significant agreement between themselves (60%), or with the PR criterion (40% and 70%), or the TAT ratio (50% and 60%). The lack of agreement with the latter score may be either a function of the unreliability of the judgments or a basic lack of comparability between the global and more atomistic methods involved.

Using the Rorschach, the agreement between judges was somewhat better (70%) but still short of significance. Agreement between both judges and PR was 50% or exactly chance probability. Agreement between both judges and the dichotomous classification based on the Rorschach Dependency percentage score was 80%. The probability of either of these judgments occurring by chance was less than .02. It would appear from this analysis that global judgments based on the Rorschach were related to the scores derived from the more atomistic scoring method, but were no more valid than those scores.

Factor Analysis of the Combination Scores

The matrix contained in Table 1 was factor analyzed using a principal components

program on the IBM 650. One large factor accounted for 92% of the variance contained in the communality estimates derived from multiple correlations on each variable.³ The significant loadings (significance arbitrarily set at .30) on this General Dependency factor were as follows: PRs (.52), SRs (.59), Q (.61), EPPS (.68), and SC (.46). TAT and Rorschach had loadings of negligible magnitude (−.02 and −.23). The ACE had a significant negative loading (−.45).

The conclusions are essentially those drawn from the matrix. The three more direct tests demonstrate the greater validity, the SC test is intermediate, and the two most indirect tests demonstrate no validity. The loading of intelligence (ACE) could be interpreted in one of two ways: (a) intelligence is an extrinsic factor influencing the measures through response sets which are also extrinsic to the trait dependency, (b) intelligence is a factor which is intrinsically related to dependency and its evaluation in others. Although Hypothesis *a* would explain the correlation between ACE and the peer ratings (see Table 1), it is hard put to explain the correlation between ACE and EPPS, a test carefully controlled for the usual response sets, or the correlation with SC, a free-response test. It must be remembered that a high intelligence level gives one a greater potentiality for independence particularly in the academic and work situations.

Factor Analysis of the Individual Scores

All of the single variable scores were inter-correlated, and the resulting 23×23 matrix was factor analyzed using Indiana University's IBM 650. Five factors were found to account for 97% of the variance contained in the estimated communalities.⁴ These factors

³ A copy of this factor matrix has been deposited with the American Documentation Institute, Order Document No. 6756 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance \$1.25 for microfilm or \$1.25 for photocopies. Make checks payable to: Chief, Photoduplication Service, Library of Congress.

⁴ Copies of the correlation matrix and the rotated factor matrix have been deposited with the American Documentation Institute, Order Document No. 6756 from ADI Auxiliary Publications Project, Pho-

were rotated to an orthogonal solution using the normalized varimax IBM program.⁵

Factor I was the most general Independence-Dependence factor obtained. Loadings of over .30 and in the expected direction were found for the Sum PR, all of the SRs, the Gough Dominance scale, all of the EPPS scales (with the exception of Abasement), and Autonomy in the SC test. Loadings in the unexpected direction were obtained from Abasement in the SC and TAT tests. These two loadings are perhaps explainable by the fact that the total score of all five variables was not partialled out of the scores for single variables. These total scores (measures of the intensity or scorability of the free-responses) were found to correlate negatively with many of the dependency measures, and high and positively with the Abasement variable in both tests.

As in the previous factor analysis, the more direct measures seemed to have the highest loadings on the most general dependency factor.

Factor II might be labeled Dominance vs. Abasement. It had positive loadings from the Gough Dominance and EPPS Dominance scales, and negative loadings from the Navran scale, EPPS Abasement, and SC Abasement scales. An examination of the Navran scale reveals that most of the items do have an abasement type of content with a lesser number of succorance items. EPPS Succorance also had a negative loading on this factor. The factor seems to involve assertion, leadership, and self-confidence at one pole; and yielding, resignation, inferiority feelings, and feelings of depression and fear at the other. The factor might relate to Cattell's (1957) Dominance vs. Submission factor, but it also seems to involve an emotionality element.

Factor III might be labeled Autonomy vs. Deference. It had positive loadings from the

Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance \$1.25 for microfilm or \$1.25 for photocopies. Make checks payable to: Chief, Photoduplication Service, Library of Congress.

⁵ Users of IBM rotation programs may be interested to know that the quartimax program was also used in order to compare it with the varimax program. Both programs yielded substantially similar results, and differences in loadings were minimal.

SC Deference and TAT Deference scores, and negative loadings from the SC Autonomy and TAT Autonomy scores. In addition, the TAT Succorance score had a negative loading and the ACE loaded positively on the factor. This factor is interesting in that it is derived entirely from two projective, or free-response, techniques, and has some obvious construct consistency.

Factor IV might be labeled Succorance. It had positive loadings from the Navran scale, EPPS Succorance, and SC Succorance. A negative loading was obtained from EPPS Autonomy. A puzzling positive loading was obtained from TAT Dominance.

Factor V was practically uninterpretable. It had negative loadings from SC Dominance and Autonomy on the one hand and TAT Abasement and Rorschach Dependency % on the other. In a subsequent analysis, this factor was collapsed and its variance distributed among the first four factors. This operation did not change the composition of these first four factors.

Beyond Factor I, General Dependency, the factors did not show extensive breadth across tests. Construct validity seems to span three tests at the most. Usually, tests which are closer in their degree of directness share some common variance on comparable variables.

The last factor analysis also exposed some of the obvious weaknesses of the tests. Although all three of the self-ratings loaded on the general dependency factor they did not discriminate among the other dimensions of the construct. Although the Navran scale is called "Dependency" it seems to be more of an abasement scale. The Rorschach had only one loading of any magnitude, and this was on a weak and uninterpretable factor. Proponents of the Rorschach might take some heart from an interesting side-finding in an unpublished study by Levitt and Zuckerman comparing the characteristics of Volunteers and Nonvolunteers for a hypnosis experiment, and using this sample of subjects. The Rorschach Dependency score was the only one which distinguished these two groups with the Volunteers scoring higher on it ($p < .01$).

But in general, whether we consider concurrent validity (correlation with the peer rating) or construct validity (loading on a gen-

eral dependency factor), the more direct tests seem to do better than the less direct tests. There is a certain irony in these results if one considers the time and effort expended on them. The self-ratings required only a few minutes to administer, and a technician could record one subject's score in less than a minute. The TAT took about an hour to administer and about an hour each for three trained psychologists to score one subject's record. If one starts calculating cost in terms of return (validity) a kind of moral can be drawn: try out simple techniques before resorting to complex techniques. This is not the first study to suggest that empirically developed, objective, and simple psychometric instruments may do better, just as well, or just as poorly as more complex free response techniques.

There are, however, important reservations to forming sweeping conclusions from the results of this study:

1. The concurrent validity of self-ratings may be a function of their greater similarity in form to the peer ratings. However, one would be rather foolish not to make self-ratings as congruent as possible with the criterion being predicted.

2. The peer and self-rating methods are lacking in that they do not discriminate dimensions within the construct, and are vulnerable to halo effect.

3. The results may not have applicability for other kinds of constructs. Simple techniques such as the Taylor scale (1953) and an Affect Adjective Check List devised by the senior author (Zuckerman, 1960) seem to work well in measuring "anxiety," but one wonders what the results would be with a less socially acceptable symptom like "hostility." One would expect that the less socially acceptable constructs would be more difficult to measure with direct techniques, but it does not follow from this that indirect techniques would do better.

4. The results may be in some part a function of the particular type of subjects used. As a group, the Indiana student nurse seems to be more abasing and nurturant, and less dominant than Edwards' (1954) college females on the EPPS. Perhaps dependency (particularly the submissive component) is

more socially acceptable in terms of the "nurse" ego-ideal. This particular group of nurses is also brighter than average, and higher on the EPPS Intraception score. Direct techniques may be less useful with duller or less "self-analytic" groups.

Allport (1953) has suggested that direct measures of motivation will be more effective than indirect measures with normal subjects, while indirect, or projective, techniques may be useful in assessing the maladjusted. A conservative conclusion from this study would be a simple underlining of Allport's (1953) statement: "A psychodiagnostician never should employ projective methods in the study of motivation without at the same time employing direct methods" (p. 111).

SUMMARY

The purpose of the study was to test the validity of a range of direct and indirect tests against a peer rating criterion (concurrent validity) and the factors derived from a factor analysis of all measures (construct validity). Another purpose was to clarify the dimensions of the broad construct, "dependency."

The subjects were 72 student nurses. They rated themselves (self-rating) and each other (peer rating) on three bipolar, eight-point scales derived from Leary's Interpersonal Adjective Check List. All subjects took the Gough Dominance and Navran Dependency questionnaires, the Edwards Personal Preference Schedule, the Rohde Sentence Completion Test, and a group administered TAT and Rorschach. The EPPS, SC, and TAT tests were scored for five relevant Murray needs: Autonomy, Dominance, Succorance, Abasement, and Deference. The Rorschach content was scored using a system adapted from DeVos. Combination scores were obtained by: summing the three peer ratings, summing the three self-ratings, combining the two true-false questionnaires, and using ratios combining the five Murray needs on the EPPS, SC, and TAT. Combination scores were correlated with peer ratings and with each other, and the resulting matrix was factor analyzed. Individual scores on all tests were analyzed in a second factor analysis.

Using combination scores, the self-ratings, questionnaires, and EPPS scores correlated

significantly with the peer ratings, while the SC, TAT, and Rorschach did not. The magnitude of the validity correlations tended to drop as function of the indirectness of the tests. More global judgments of the TAT and Rorschach, using extreme groups on the peer rating distribution, did not indicate any greater concurrent validity for these tests. A factor analysis of the combination scores yielded one large factor called General Dependency. All of the four most direct tests showed moderate to high loadings on this factor while the less direct tests (TAT and Rorschach) showed negligible loadings. A factor analysis of the individual scores yielded four interpretable factors: (a) General Dependency, (b) Dominance vs. Abasement, (c) Autonomy vs. Deference, (d) Succorance. Factors a, b, and d were composed mainly of individual scores from the peer and self-ratings, questionnaires, EPPS, and SC tests. Factor c was composed entirely of scores from the SC and TAT tests.

In general, the more direct measures of dependency demonstrated the greater validity of both types, but the conclusions about the relative validity of the two types of tests are limited by the form of the measures, the particular type of subjects used, and the particular construct investigated.

REFERENCES

- ALLPORT, G. W. The trend in motivational theory. *Amer. J. Orthopsychiat.*, 1953, 23, 107-119.
- CATTELL, R. B. *Handbook for the Sixteen Personality Factor Questionnaire*. Chicago: Psychometric Affiliates, 1957.
- DEVOS, G. A quantitative approach to affective symbolism in Rorschach responses. *J. proj. Tech.*, 1952, 16, 133-150.
- EDWARDS, A. L. *Edwards Personal Preference Schedule: Manual*. New York: Psychological Corporation, 1954.
- GOUGH, H. G., MCCLOSKEY, H., & MEEHL, P. E. A personality scale for dominance. *J. abnorm. soc. Psychol.*, 1951, 46, 360-366.
- HORNEY, K. *Our inner conflicts*. New York: Norton, 1945.
- HORST, P. A generalized expression for the reliability of measures. *Psychometrika*, 1949, 14, 21-23.
- LEARY, T. *Interpersonal diagnosis of personality*. New York: Ronald, 1957.
- MURRAY, H. A. *Explorations in personality*. New York: Oxford Univer. Press, 1938.

- NAVLAN, L. A rationality derived MMPI scale to measure dependence. *J. consult. Psychol.*, 1954, **18**, 192.
- ROHDE, AMANDA R. *The sentence completion method*. New York: Ronald, 1957.
- TAYLOR, JANET A. A personality scale of manifest anxiety. *J. abnorm. soc. Psychol.*, 1953, **48**, 285-290.
- ULLMAN, L. On the relationship between amount of hospitalization and self-assertion. *Amer. Psychologist*, 1958, **13**, 327. (Abstract)
- ZUCKERMAN, M. The validity of the Edwards Personal Preference Schedule in the measurement of dependency-rebelliousness. *J. clin. Psychol.*, 1958, **14**, 379-382.
- ZUCKERMAN, M. The development of an Affect Adjective Check List for the measurement of anxiety. *J. consult. Psychol.*, 1960, **24**, 457-462.
- ZUCKERMAN, M., & GROSZ, H. J. Suggestibility and dependency. *J. consult. Psychol.*, 1958, **22**, 328.
- ZUCKERMAN, M., NORTON, J., & SPRAGUE, D. Acquiescence and extremes sets and their role in tests of authoritarianism and parental attitudes. *Psychiat. res. Rep.*, 1958, **10**, 28-45.

(Received June 20, 1960)

RESPONSE BIAS IN QUESTIONNAIRE REPORTS¹

N. H. AZRIN, W. HOLZ, AND I. GOLDDIAMOND²

Anna State Hospital, Illinois

If one defines psychology as the study of behavior, the direct measurement of behavior appears to be a minimal prerequisite to further analysis. Several alternatives to a direct measurement of behavior are commonly practiced. One such indirect method defines and measures the behavior in terms of the effect of that behavior upon the environment. For example, the measurement of reaction time typically is based upon the moment of closure of an electric switch or push-button. The simple fact of closure of the push-button does not guarantee that the movement of any one finger be involved since any finger or even the palm, wrist, arm, or leg might just as easily have been used. Such ambiguities in interpretation are easily overcome, and the experimenter can easily confirm his interpretation of the switch closure by occasionally or continuously observing the behavior directly.

A second alternative to direct behavioral observation is the interview or questionnaire procedure. Here the experimenter typically does not have any simple means of direct behavioral observation. Rather, the subject himself is expected to observe his own behavior and to describe it at some future date. The subject's reports usually cannot be evaluated by direct observation of the behavior being reported as was true of closing of the switch. The problem is often enhanced by the fact that the behavior being reported upon is basically unobservable by the experimenter by its very nature. This is true for the so-called "subjective reactions" as when an individual states that he feels hostile or afraid. In addition, it is quite likely that a report of one's own behavior will be modified consid-

erably by the audience or experimenter to whom the report is being made. Other factors such as social acceptability may also be involved. The reply to the question "Did you cheat?" would probably be different if the interviewer were a classmate than if he were an instructor. For whatever the ultimate reasons may be, the individual being questioned may have a pre-existing tendency or bias in admitting to some statements and not to others. The present study was performed in order to study the influence of such response biases upon the reports of behavior obtained through a questionnaire.

A well-known study by Shaffer (1947) deals with the reports of combat flyers of their fears in combat. The reports had been obtained from these flyers by means of a questionnaire some 2 months following the termination of their combat experiences. This study has been widely interpreted as demonstrating that some behavioral reactions such as "soiling one's pants" are more indicative of fear in combat than are behavioral reactions such as "feeling nervous and tense." In order to evaluate the validity of this interpretation, 160 college freshmen and sophomores, including males and females, in five separate psychology and sociology classes were given a questionnaire. This questionnaire contained the same 15 "symptoms" reported in the original investigation by Shaffer. All of the questionnaires had the following directions:

Imagine that you are a combat flyer who has flown many missions over enemy territory. Your commanding officer gives you the questionnaire below and tells you to fill it in. Fill in the answers keeping in mind what your commanding officer expects you to have felt.

Two forms of the questionnaire were used, however; half of each class of students re-

¹ This investigation was supported by a grant from the Psychiatric Training and Research Fund of the Illinois Department of Public Welfare.

² Now at Arizona State College, Tempe, Arizona.

ceived one form and half received the other form. One form stated after the first sentence:

You have been extremely frightened on all of your missions and have experienced each of the symptoms below on every flight.

The other form stated:

You have never been frightened on any of your missions and have not experienced each of the symptoms below on every flight.

One half of the students are thereby told that they have never experienced any of the listed symptoms, whereas the other half are told they have experienced all of the symptoms. Further, all students were instructed by the questionnaire to answer in terms of what is expected, regardless of what behavior is presumed to have occurred.

RESULTS AND DISCUSSION

Table 1 presents the percentage of students who chose each symptom as occurring "often" or "sometimes." It will be recalled that the students had been told that all symptoms were experienced equally often. The only basis for checking some symptoms more than others was the specific instruction to keep in mind what answers are expected. Had there been no predisposition or response bias toward some symptoms, one should expect all symptoms to have been selected equally often. Certainly, no particular rank ordering of the symptoms should have emerged. The results of Table 1 demonstrate that the selection of symptoms does not follow a random distribution. Some symptoms were selected as much as six times as often as others.

Spearman rank-order correlations were performed to determine the consistency of this response bias among the students. It was found that the rank-order of symptoms for males correlated with that of females with $\rho = .88$. Similarly, the rank-order correlation of symptoms was .95 between those students who were told that they had experienced all the symptoms and those who were told that they had not experienced the symptoms. This high degree of similarity demonstrates that the response bias toward certain symptoms exists regardless of whether or not the symptom was alleged to occur.

TABLE 1
REPORTED SYMPTOMS OF COMBAT FEAR OF
160 COLLEGE STUDENTS

During combat missions did you feel:	% of students stating "often" or "sometimes"
That your muscles were very tense	72
A pounding heart and rapid pulse	71
"Butterflies" in the stomach	67
Dryness of the throat or mouth	67
"Nervous perspiration" or "cold sweat"	61
Sense of unreality that this couldn't be happening to you	49
Easily irritated, angry, or "sore"	43
Need to urinate very frequently	42
Trembling	39
Unable to concentrate	36
Sick to the stomach	34
Right after a mission, unable to remember details of what happened	32
Confused or rattled	28
Weak or faint	25
That you have wet or soiled your pants	11

In order to determine whether the same response bias might have affected the reports of the combat flyers, a Spearman rank-order correlation was performed between the symptoms reported by the flyers and those reported by the students. It was found that the rank orders of the responses were highly similar ($\rho = .89$) between the students and the flyers. Nor was this relationship reduced for those students who were told that they had not experienced the symptoms ($\rho = .94$) as compared with those who were told that they had experienced all of the symptoms ($\rho = .90$). The statistical stability of this response bias is evidenced by the degree to which the rank-order in each classroom of students was correlated with the rank-order reported by the combat flyers: $\rho = .70, .82, .85, .90$, and $.92$.

The response pattern obtained from the students by means of the questionnaire is al-

most completely predictable on the basis of response bias. Therefore, it is quite likely that the same type of response bias operated on the combat flyers. Any conclusions concerning the actual symptoms must await study by a method that provides for a more direct and objective measurement.

The present findings may well be considered for their implications for the use of interview and questionnaire methods in general. Unless an objective and direct means of measurement is available, the questionnaire re-

sponses may be independent of the behavior being studied. A definitive method of determining the validity of the reports is the direct and objective measurement of the behavior being reported. Once such a direct measure is available, however, the very need for questionnaire reports is eliminated.

REFERENCE

- SHAFFER, L. F. Fear and courage in aerial combat. *J. consult. Psychol.*, 1947, 11, 137-143.

(Received June 21, 1960)

THE MMPI AS A MEASURE OF CHARACTER STRUCTURE AS REVEALED BY FACTOR ANALYSIS¹

JOSEPH C. FINNEY

Hawaii Department of Health, Honolulu

The psychologist who needs conceptual tools to understand and to help people, faces a dilemma. On the one hand he can accept and use dynamic concepts, usually psychoanalytic ones, including urges and defense mechanisms, which seem fairly satisfactory in working with patients. If he does so, it is with the disadvantage that he assumes some cause and effect relationships on slender evidence, and that he treats as quantities many things that he has no way of measuring. On the other hand, if he is hard-boiled, skeptical, operational, and parsimonious, he may confine himself to relationships experimentally or statistically established, and to quantities that can be accurately measured. If so, he hampers himself more seriously: his concepts are at a level too crude and simple to be of value in grasping and dealing with his patients' troubles.

There is, further, a variety of opinion about what quantities an objective psychological test should measure (Cattell, 1957a; Gough, 1957; Hathaway, 1951). Even factor analytic studies (Cattell, 1957b; Cook & Wherry, 1950; Cottle, 1950; Lingo, 1960; Welsh, 1956; Wheeler, Little, & Lehner, 1951; Wil-

¹ The following materials have been deposited with the American Documentation Institute: Tables A through M, significant reference vector loadings, for factors comparable among groups; lists of items in new scales; intercorrelations of the 59 variables; successive latent roots of factors; complete tables of loadings of the first two centroid factors; *D* and inverse of *D*; correlations among oblique factors; complete tables of reference vector and primary factor loadings of oblique factors. Order Document No. 6757 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance \$2.75 for microfilm or \$7.50 for photocopies. Make checks payable to: Chief, Photoduplication Service, Library of Congress.

liams & Lawrence, 1954) have not agreed on what fundamental dimensions are uncovered. It seems desirable that tests aimed at variables of importance in clinical, including psychoanalytic work, be included in factor analytic studies. The recent development of programs by which factor analysis, including rotation to oblique simple structure, can be performed entirely by electronic computer, (Dickman, 1959; Pinzka & Saunders, 1954) makes it possible for the psychologist with minimal mathematical training and little time for computation to do such studies.

AIMS OF THIS STUDY

The battery of measurements used in this study evolved gradually from the clinical demands of a county mental health clinic. For nearly 4 years the author and other psychologists made blind interpretations from the MMPI scores of applicants for clinic services. These personality descriptions were compared with the reports of interviews, and attempts made to account for discrepancies. One result was to increase the psychologists' skill in drawing inferences from MMPI scores. Even so, there remained many areas, regarded as important in clinical interviews, which no standard MMPI scale seemed to measure. One by one, additional scales were added to fill the gaps.

The strongly felt need for a measure of strength of ego-ideal or conscience was met by Gough's (1957) responsibility scale. The clinical desire for measures of hysterical character (defense of repression) and paranoid character (defense of projection) was tentatively met by adding to the battery Wiener's (1948) "subtle" subscales of *Hy* and *Pa*, which sounded promising. The "ego functions" were

tapped by some of Gough's (1957) scales, such as *Ie* (intellectual efficiency) and *Do* (dominance). Finally, some still untouched areas, such as stinginess and warmth, were included by adding several unpublished experimental scales devised for the purpose by the author. With the aid of the supplementary scales, the psychologist interpreting blindly from test scores became able to deal with the aspects of personality that interested the clinical interviewers, and with a satisfactory degree of agreement.

The question arises, which of the many measures represent important basic aspects of personality? Cattell (1952) has contended that factor analysis with rotation to oblique simple structure eliminates the superficialities and reveals the fundamental underlying dimensions. It seemed appropriate to apply that method to the group of personality measures that were used.

METHOD

A questionnaire of 600 items was given routinely to all willing literate, nonpsychotic adults applying for services (whether as patients or as parents of child patients) at a community mental health clinic in Illinois, until 50 men's and 50 women's records were obtained (a process which took about 6 months). The questionnaire consisted of the booklet form of the MMPI, with some unscored and duplicating items omitted and some additional items included.

Tests were scored for 56 scales and two individual items. Intercorrelations were computed for 59 variables (the 58 test variables plus sex), using raw scores. The variables used were:

1. Sex (male or female)
2. Sex experience (A single item: "I've had sex relations with at least three different people in my life.")
3. Anti-Catholic prejudice (A single item: "I have no religious prejudice, but I think it would be bad for the country if a Catholic were elected President." The study was done in 1959.)
4. *L*
5. *F*
6. *K*
7. *Hs* (without *K* correction)
8. *HsK* (*Hs* with *K* correction)
9. *D*
10. *Hy*
11. *Pd* (without *K* correction)
12. *PdK*
13. *Mf*
14. *F* minus *K* (Gough, 1950)
15. *PaO* (Wiener's, 1948, "obvious" subscale of *Pa*)

16. *Pa*
17. *Pl* (without *K* correction)
18. *PlK*
19. *Sc* (without *K* correction)
20. *ScK*
21. *Ma* (without *K* correction)
22. *MaK*
23. *Si* (Drake, 1946)
24. *A* (Welsh, 1956)
25. *MA* (manifest anxiety, Taylor, 1953)
26. *Ie* (intellectual efficiency, Gough, 1957, MMPI items only)
27. *St* (capacity for social status, Gough, 1957, MMPI items only)
28. *Es* (ego strength, Barron, 1953)
29. *Do* (dominance, Gough, 1957, MMPI items only)
30. *Re* (responsibility, Gough, 1957, MMPI items only)
31. *H₁* (an experimental scale on hostility, 117 items)
32. *H₂* (an experimental scale on hostility, 146 items)
33. *Sc* minus *Ma*
34. *W₂* (an experimental scale on warmth, 81 items)
35. *ODy* (an experimental scale on optimistic dependency, 9 items)
36. *Dy* (dependency, Navran, 1952)
37. *Ph* (an experimental scale for phobia, 12 items)
38. *Sd* (an experimental scale for sadism, 12 items)
39. *Ob* (an experimental scale for obsession, 16 items)
40. *An* (an experimental scale for anal or compulsive character, containing Scales 41, 43, 44, and a few additional items, 48 items)
41. *Rig* (rigidity-flexibility, Gough, 1957, reduced to 15 items)
42. *AuF* (the Berkeley F Scale, Adorno, Frenkel-Brunswick, Levinson, & Sanford, 1950, reduced to 15 items)
43. *Or* (an experimental scale for orderliness, 9 items such as: "I like to have my clothes clean at all times.")
44. *Sti* (an experimental scale for stinginess, 6 items such as: "I have to admit I hate to see money wasted.")
45. *Stb* (an experimental scale for stubbornness, 10 items such as: "People can push me just so far, and then I have to take a stand.")
46. *Reb* (an experimental scale for rebelliousness, 20 items)
47. *Sub* (an experimental scale for submissiveness, 26 items)
48. *Ap* (accepted passivity, Harris, unpublished)
49. *Id* (an experimental scale intended to measure inner direction versus other direction—concept of Riesman, Glazer, & Denney—21 items)
50. *De* (delinquency, Gough & Peterson, 1952, re-named socialization, Gough, 1957, MMPI items only)
51. *Fm* (feminine masochism, Hecht, 1952)
52. *PaS* (Wiener's, 1948, subtle subscale of *Pa*)
53. *R* (Welsh, 1956)

54. *HyS* (Wiener's, 1948, subtle subscale of *Hy*. Our selection followed a somatic-nonsomatic division, and differed slightly from Wiener's in including Items 137 and 179 and excluding Item 190. Little's *Dn* scale (Little & Fisher, 1958) is also virtually identical with *HyS*)

55. *Rep* (an experimental scale for repression, containing the 29 *HyS* items plus 17 others)

56. *Em* (an experimental scale for embarrassment, 13 items)

57. *Fe* (femininity, Gough, 1957, MMPI items only)

58. *Sc* minus *Pt*

59. *Hy* minus *PtK*

A Thurstone centroid factor analysis was performed by electronic computer. Rotation to oblique simple structure was also performed entirely by electronic digital computer, using a standard program (Dickman, 1959; Pinzka & Saunders, 1954) so that no human judgment entered into the selection of the angles of rotation. Both reference vector loadings and primary factor loadings were obtained.

Intercorrelations and factor analyses were performed separately on the group of 50 men (hereinafter called Group M), the group of 50 women (called Group W), and the combined group (Group C).

A question may be raised about the use of scales with common items. The standard MMPI scales have many items common to three or more scales, and the subscales used here compound this practice. The entire *PaS* scale of Wiener, for instance, is included in the *Pa* scale. Welsh (1956) has considered this problem serious enough to eliminate all common items before calculation. Any correlation, however, may be considered as the result of hypothetical common elements (McNemar, 1949, pp. 117-118), and it has been pointed out (Wheeler, 1951) that the problem of correlations among scales is the same whether it is due to physically identical items or to common meanings in different items.

RESULTS AND DISCUSSION

Centroid Factors

Using the criterion of continuing to extract factors as long as the latent root exceeds 1.0, 13 factors were found in Group M, 12 in Group W, and 12 in Group C. The 12 to 13 factors extracted accounted for 86.4% of the variance in Group M, 85.3% in Group W, and 82.0% in Group C.

The Thurstone centroid factor analysis produced, in all three groups, a first factor most heavily loaded on uncorrected *Pt* (reference vector loadings .90, .95, .94, dropping to .73, .86, and .85 with *K* correction), Welsh's *A* (.87, .92, .91), Navran's *Dy* (.89, .92, .92), uncorrected *Sc* (.86, .93, .90), *F*-minus-*K*

(.75, .89, .81), and Taylor's *MA* (.85, .89, .89); and with substantial negative loadings on Gough's *Ie* (-.76, -.85, -.85) and Barron's *Es* (-.78, -.85, -.85) (Table A).² This factor accounted for 28.7% of the variance in Group M, 36.6% of the variance in Group W, and 33.3% of the variance in Group C.

In all three groups, the second factor was most heavily loaded in the positive direction with Welsh's *R* (.45, .60, .68), Harris' *Ap* (.40, .47, .53), Gough's *Re* (.53, .32, .62), Drake's *Si* (.53, .33, .38), and experimental scales for embarrassment and submissiveness; and loaded negatively with *Ma* (-.66, -.51, -.61) and with experimental scales for stubbornness and rebellion (Table B). The first two factors together accounted for 40.5% of the variance in Group M, 45.6% of the variance in Group W, and 44.2% of the variance in Group C.

The first two factors obtained by the Thurstone centroid method correspond closely with those previously obtained by Welsh (1956) using a similar procedure. Welsh's scales *A* and *R* (1956) were designed to measure these dimensions.

Thurstone centroid factor analysis defines dimensions with the maximum common variance, i.e., in which the greatest number of measures will vary together, or in agreement with each other. That procedure seems designed to maximize the influence of "response set." Examination of the evidence suggests that this is indeed the case, and that Welsh's (and the present) first two centroid factors, *A* and *R*, are very largely composed of response set.

One much studied response set is the tendency to present oneself as, on the one hand, psychologically healthy and normal, or, on the other hand, as emotionally upset and in need of help. This aspect has been described by the Minnesota group as "faking good" or "faking bad," by Gough (1950) as "dissimulation," by Edwards (1957) as "social desirability," and by DeSoto and Kuethe (1959) as "the set to claim undesirable symptoms." This falsifying tendency, whether conscious or not, is

² Tables A through M are included in the materials available through the American Documentation Institute.

believed to influence questionnaire test scores seriously, the more so in tests with much face validity or obviousness. Repeated efforts have been made to eliminate this source of variance from clinical tests, either at the source (Edwards) or by suppressor variables (the earliest of such efforts being Meehl's *K* correction—Meehl & Hathaway, 1946).

Evidence suggests that the first centroid factor consists largely of this response tendency, the willingness or unwillingness to admit psychological sickness. The evidence consists of the high loadings on this factor of several scales already known to measure this tendency. One such scale is Gough's (1950) *F*-minus-*K* dissimulation index, with reference vector loadings of .75, .89, and .81 on this factor. Another is uncorrected *Pt*, much of whose variance DeSoto and Kuehe (1959) have shown to represent a "set to claim undesirable symptoms," and which has loadings of .90, .95, and .94 on this factor.

Taylor's and Welsh's "anxiety" scales are very highly correlated with uncorrected *Pt* (e.g., in Group C, .92 and .93, respectively), and may also be described as consisting largely of this response set. Both Welsh's and Taylor's scales consist of items recognizably "sick" in their frank verbal content—items that might easily be avoided by a person seeking to appear "normal." With these considerations, it is easy to understand the failure of Taylor's scale (Kendall, 1954; Lauterbach, 1958) to be validated as a measure of anxiety.

It is noted that *K* correction, which increases the validity of five clinical scales (Meehl & Hathaway, 1946), invariably decreased their loadings on this factor. If a factor measures genuine pathology, the suppressor variable *K* should increase factor loadings.

The second centroid factor showed no consistent relationship to measures of dissimulation. This is consistent with Wiggins and Rumrill's (1959) finding of lack of relationship between dissimulation and Welsh's *R* scale.

A second response set that has been widely studied is the tendency to answer "true" or "false." (A good review of the literature on response sets has been done by Wiggins—1959; Wiggins & Rumrill, 1959—and will not

be repeated here.) There are some evidences to suggest that the second centroid factor consists largely of this "response bias."

Welsh derived his *R* scale (1956) to measure his second centroid factor. The present study replicates Welsh's in finding substantial loadings (.45, .60, .68) of the second centroid factor on the *R* scale. All 40 items of the *R* scale are keyed false.

Couch and Keniston (1960) investigated the True or False response tendency in detail, and described the personality characteristics of "yeasayers" and "naysayers" in terms strikingly like those used by clinicians for low and high *R* scorers. Couch's "overall agreement score," however, gives evidence of representing not only yeasaying but also the other type of response set; its correlations with MMPI variables and with 16 PF resemble the second order factor that Cattell calls "anxiety" (perhaps better named "willingness to admit sickness") and Welsh's and the present first centroid factor. This is not surprising, since in most questionnaires the Yes answers tend to be the sick sounding ones. The present study may have succeeded in separating these tendencies only in an orthogonal factor analysis.

As a check on yeasaying and naysaying, the loadings of the scales on the second centroid factor, for Group C, were correlated with the percentage of true keyed items in the scales. Single-item variables as well as *Mf* (spurious for mixed-sex groups because of scoring differences) were eliminated. The result is only suggestive, since scales contained common items, but the correlation of $-.60$ is at least consistent with the view that the second centroid factor is related to a Naysaying response bias. (The first centroid correlated .37 with "percentage true," no doubt a result of the aforementioned tendency of True answers to be sick sounding.)

To identify the first two centroid factors (and Welsh's *A* and *R* scales) as largely response tendency is not to discard them as trivial. The clinician values a measure of the first factor, willingness or unwillingness to admit psychological sickness, as a sign of a patient's motivation for treatment. And Couch and Keniston (1960) have shown personality correlates (id and impulsiveness versus super-

ego and inhibition) for the second response set.

Obliquely Rotated Factors, Cross-Validated

General. Both of the response set factors disappeared on rotation to oblique simple structure, and a quite different set of factors appeared. Indeed, some of the scales most heavily loaded with the first centroid factor, such as Welsh's and Taylor's anxiety and *F*-minus-*K*, were remarkable for showing little or no loading on any of the oblique factors.

This result is to be expected. A response set, almost by definition, influences many scales. The centroid method, as pointed out above, is designed to maximize the influence of response sets. Rotation to oblique simple structure, however, is designed otherwise. The theory behind this procedure assumes that any one scale should be related to only a very few fundamental dimensions, and hence this method seeks factors each having near-naught loadings on as many scales as possible. A vector such as the first centroid, which has substantial loadings on most of the scales used (Table A) is avoided. It should be impossible for this procedure to define a factor substantially loaded with a widespread response set. Cattell (1952) has given reasons for thinking that rotation to oblique simple structure brushes the superficialities away and discovers the fundamental underlying dimensions.³

The rotation was performed exclusively by electronic digital computer in order to avoid the possibility that the investigator's preconceived notions of personality dimensions might influence the type of factors found.

Examination of the 12 to 13 factors extracted from each group showed that 5 or possibly 6 factors were common to all three samples. These are shown in Tables C, D, E, F, G, and H. The factors that were found in both Groups M and W (and also in the com-

bined Group C) can be regarded as confirmed by cross-validation. The factors found in only one group may be regarded as having failed in cross-validation, though some of them may be genuine factors peculiar to one sex. Only a further cross-validation on a new sample can decide this question.

Factor 1: Anal compulsive character or reaction formation. One factor corresponded closely with Freud's 1908 description of the anal character (Freud, 1950). It contained substantial loadings on three experimental scales designed to measure Freud's "anal triad": orderliness (.61, .61, .61), stinginess (.34, .23, .41), and stubbornness (.40, .21, .30), respectively (Table C). These three scales contained no common items. Higher than any of these were the loadings on Gough's (1957) rigidity-flexibility scale (.70, .74, .70). A still higher loading (.73, .71, .78) was on an experimental scale designed to measure anal character, containing the four scales just mentioned.

An abbreviated form of the Berkeley authoritarian F Scale (AuF) also showed positive loadings on this factor (.32, .15, .24).

This factor seems clearly to correspond to the personality type known commonly as the anal or compulsive character. In addition to Freud's original paper in 1908, this type has been described by Jones (1913) and by Fenichel (1945). The concept of compulsive character has had widespread clinical acceptance and is recognized as an official diagnosis by the American Medical Association and the American Psychiatric Association (1952).

Despite clinical acceptance, there has hitherto been scant experimental evidence to support the concept of anal or compulsive character. In the one outstanding positive study, Sears (1942), dealing with ratings of fraternity brothers by each other, found (when halo effect was partialled out) positive correlations of .36-.39 among ratings of orderliness, stinginess, and stubbornness. His finding was all the more remarkable because the young men regarded orderliness as desirable and stinginess and stubbornness as undesirable.

The present finding of a strong relationship between the anal personality factor and Gough's rigidity scale is in accord with the

³ It should be noted that this procedure does not eliminate response set from the items themselves. Hence an item analysis, using an oblique factor as criterion, may still produce a scale biased with response set, unless special precautions are taken. This is evidently the case with Cattell's 16 PF, in which 7 of the 16 scales are correlated both with *F* and also (in the reverse direction) with *K*, evidence of dissimulation effect (Karson & Pool, 1957).

common clinical impression of the compulsive person as "rigid" and suggests that the relationship between this factor and the experimental findings on perceptual rigidity (Luchins, 1942; Rokeach, 1948) should be explored. Since the experimental scales *Or*, *Sti*, and *Stb* have not been separately validated, Gough's rigidity scale must be regarded as the chief defining scale for this factor. It is noted that the rigidity scale has higher loadings than orderliness, stinginess, and stubbornness, and it is suggested that rigidity is the most central characteristic of the anal or compulsive factor.

Anal character was the factor that accounted for the largest portion of the total variance. The remaining factors are presented in arbitrary order. Wherever necessary, a factor from one of the patient groups has been reversed in sign, in all its loadings, to make it comparable with a factor derived from another patient group.

Factor 2: Hysterical character or repression. Another factor common to the three samples was one whose highest loading (among previously published scales) was on Wiener's subtle *Hy* scale or *HyS* (.58, .49, .58). It also had smaller positive loadings on *Hy* (.39, .16, .28) and *K* (.15, .13, .26), and negative loadings on *F* (−.18, −.33, −.22) (Table D). The highest loading (.69, .78, .70) was on *Rep*, an experimental scale designed to measure repression or hysterical character, and containing all of the *HyS* items with others.

This factor also seems to correspond to a character type long accepted in psychoanalysis (Fenichel, 1945). Rosenzweig (1945) described an "impunitive" reaction characteristic of hysterical persons, and a projective test for measuring it. McKinley and Hathaway (1944) found that persons with diagnosed conversion hysteria (physical symptoms such as headache and vomiting) answered certain questionnaire items in a naive Pollyanna-like manner. Wiener (1948) separated these items (*HyS*) from the full *Hy* scale. Eriksen, in a series of studies, related psychiatric diagnoses of hysteria, an MMPI measure (*Hy* minus *Pt* with *K* correction) and a perceptual measure of repression (Eriksen, 1952, 1954; Lazarus, Eriksen, & Fonda, 1951).

While hysterical conversion reaction is an official medical diagnosis (American Psychi-

atric Association, 1952), hysterical personality has never attained that status, probably because it is so common and is not considered sick enough.

This was the only oblique factor in which a dissimulation measure, *F*-minus-*K*, had a substantial loading (−.21, −.28, −.30), and it was in the paradoxical direction, such that faking good would appear to make one's score sicker. Other measures of dissimulation (*A*, Taylor *MA*, *Pt*) were unrelated to the factor. Consideration of the theory of repression suggests that it is genuine repression that is producing the appearance of dissimulation, rather than dissimulation contaminating the definition of the repression factor.

R. B. Cattell, who has kindly examined the present study, suggests (personal communication, 1960) that the hysterical character or repression factor found here may be the same as Cattell's Factor I, called "premsia." Factor II of Lingoes (1960) also resembles the present factor of repression or hysterical personality.

Each of the factors discussed so far, compulsive character and hysterical character, has a heavy loading on a previously published scale (Gough's rigidity and Wiener's *HyS*). In each case, there is an experimental scale (*An* and *Rep*, respectively) which includes the previously published scale along with other items. The *An* and *Rep* scales were attempts of the investigator to improve the measurement of the compulsive and hysterical personalities, and their higher factor loadings are evidence of a moderate degree of success in this endeavor.

One reason why *Rep* is a better measure of hysterical personality than *HyS* may be that it contains not only the *HyS* items of repression and denial, but also items designed to tap histrionic dramatization. Another reason may be its greater freedom from response sets. The *HyS* items are almost all keyed false, and, in the keyed direction, are "zero" (i.e., unpopular) responses that deny sickness and appear socially desirable (Wiener, 1948). The *Rep* scale is somewhat better balanced for both kinds of response set as well as for infrequency, and this may be why it is a purer measure of the hysterical character factor.

Factor 3: Paranoid character or projection. A third common factor was one whose highest

loading was on Wiener's subtle *Pa* or *PaS* scale (.85, .71, .76) (Table E). The whole *Pa* scale showed lower (though still substantial) loadings (.52, .34, .43) while Wiener's obvious paranoid subscale was unrelated to this factor (.06, -.10, .05). The designation "paranoid character" was applied to this dimension. This factor resembles Factor IV of Lingo (1960). Paranoid character, without psychosis, is an accepted clinical concept and is an official medical diagnosis (American Psychiatric Association, 1952).

Factor 4: Conversion. A fourth factor that was extracted separately from Groups M, W, and C, had heavy loadings on *Hs* (with and without *K* correction) (.77, .61, .59; .80, .69, .65 corrected) and on *Hy* (.44, .72, .50), and much lower loadings on *D* (.22, .20, .10) (Table F). It was clearly conversion reaction, a long accepted clinical entity, and an official medical diagnosis (American Psychiatric Association, 1952).

The first three character types found as factors, the compulsive, the hysterical, and the paranoid, have been defined by psychoanalysts (Abraham, 1953) in terms of bodily zones of libidinal fixation. Others, however, (Freud, 1946; Freud, 1936) have defined these characters in terms of the predominant defense mechanism of the ego. These characters are marked, respectively, by the defenses of reaction formation, repression, and projection. Thus, the four factors most clearly cross-validated in the present study seem to be equivalent to four defenses: reaction formation, repression, projection, and conversion.

Factor 5: Oral aggression and delinquency. A fifth cross-validated factor was marked by high loadings on *Pd* (.63, .53, .58; or .67, .62, .65 with *K* correction) and on the Gough-Petersen delinquency or socialization scale (.34, .47, .52) (Table G). It was tentatively identified as oral aggression. It is less clearly identified than other factors in terms of psychodynamics and motivation. Gough (1957) described high scorers on *De* as "demanding," "rebellious," and "exhibitionistic." Cattell (1957b) used the same three adjectives in describing one pole of his Factor D, a factor which he omitted from the 16 PF for lack of an adequate measuring scale. It seems likely that Gough's delinquency-socialization is iden-

tical with Cattell's Factor D and with the factor found in the present study.

An experimental scale on rebellion, however, did not correlate in the expected manner with this factor (.06, -.21, -.03) (Table G). In Group M, the delinquent pole of the factor is correlated .22 with stubbornness, while among the women it is correlated .23 with submissiveness. Only further study on a new group can tell whether this is a genuine sex difference, or an accidental error of sampling. It fits in, however, with the general impression that women's delinquencies tend to be more passive and less aggressive than men's.

The tentative suggestion is offered that demandingness rather than rebellion may be the central feature of this factor. If so, it can be equated with the psychoanalytic concept of "oral aggression." This factor is not interpreted as socialization, conscience, or ego-ideal because of its insignificant loadings on Gough's responsibility scale (-.02, -.23, -.04).

Factor 6. A sixth factor, one of obsessive worrying, may be common to the three groups analyzed. Each group showed a factor positively loaded with *K* corrected *Pt* (.34, .40, .52), and less with uncorrected *Pt* (.15, .29, .37). The higher loadings on corrected than uncorrected *Pt* suggest (for reasons discussed above) that this factor is a personality variable rather than a response set. Groups M, W, and C also agreed in loading *Sc*-minus-*Pt* (-.24, -.34, -.39) and *Hy*-minus-*PtK* (-.67, -.50, -.16) on the same factor. The factor patterns from Groups M and W correspond well enough with the accepted (American Psychiatric Association, 1952) clinical concept of obsessive-compulsive psychoneurosis. But the factor from Group C bearing nearest resemblance has more the flavor of general neuroticism, with additional loadings on *Hy* (.50), *Hs* (.33), *HsK* (.38) and *D* (.36).

Oblique Factors Less Clearly Established

There were several factors common to two of the three factor analyses. Group M and Group C showed a psychotic factor (Table I) with highest loadings on *Sc*-minus-*Pt*. It seems plausible that the absence of this factor in Group W was an accident of sampling.

A factor of sadism, cruelty, narrowmindedness, and prejudice appeared in Groups W and C (Table J). It was marked by loadings on the Berkeley authoritarian F Scale (abbreviated form), anti-Catholic prejudice, and an experimental scale designed to measure sadism.

Groups M and C shared a factor marked by high loadings on *Ma* (Table K). It was not significantly related to Welsh's *R*.

All three groups had factors related to masculinity-femininity (Table L), though there was no single variable with high loadings on the factors from all three groups. The variable *Mf*, which seems to do so, is artificial in that it is scored differently for men and for women. Gough's *Fe* scale had rather high loadings on the masculinity factors from Groups M and C and lower loadings on the factor derived from Group W. In Group M, the masculine pole included inner-direction, and the feminine pole, other-direction, perhaps reflecting a difference between farmers and salesmen, respectively, in this Illinois sample. In any event, there is no theoretical reason for expecting "femininity" to mean the same in men as in women.

Both Group M and Group W showed factors whose highest loading was on the single item "I have had sex relations with at least three different people in my life." There is no similarity in other respects between the factor obtained from the men and that from the women (Table M). In women, high degree of sex experience was associated with measures of warmth, of delinquency, and of accepted passivity.⁴

Three other factors appeared in Group M alone. One was positively loaded with optimistic dependency, warmth, and phobia, and

negatively loaded with *L*, *Si*, responsibility, and feminine masochism. Another consisted of (affirmatively) stubbornness, rebellion, dominance, and inner-direction, and (negatively) submission and *L*. The last comprised only delinquency affirmatively, and femininity, *R*, and stinginess in the negative direction.

Group W also showed three additional factors. One was loaded affirmatively with accepted passivity, sex experience, obsession, submission, and feminine masochism; and negatively with *Sc*-minus-*Pt*, sadism, rebellion, *ScK*, and *MaK*. It may be compared with Cattell's Factor E (1957a). Another comprised positive loadings on warmth, *PaS*, anti-Catholic prejudice, capacity for status, and phobia; and negative ones on *L*, inner-direction, *R*, *HyS*, and responsibility. The other had positive loadings on responsibility, sex experience, *L*, *Pa*, intellectual efficiency, and stinginess; and negative on phobia, femininity, and anti-Catholic prejudice.

Two other factors appeared in the combined group. One was a general hostility factor, loaded with *H*₂, *PaS*, *H*₁, stubbornness, *Pa*, and dominance. The other consisted of self-confidence and ambition, with loadings in one direction on capacity for status, intellectual efficiency, dominance, accepted passivity, responsibility, and warmth, and in the other direction for *R*, social introversion, and embarrassment.

Nature of the Oblique Factors

Some of the factors have been seen to correspond with diagnostic categories (compulsive personality, paranoid personality). But a basic conceptual difference must be pointed out. Diagnostic categories are discrete entities: a person either is or is not within the category, and the usual practice is to give a patient not more than one psychiatric label.

Factorial dimensions, on the other hand, are continuous. The question is not whether a person is or is not a paranoid character, but rather, how much paranoid character does he have? Further, a person may be high in two or more dimensions. A person high in both reaction formation and projection may, with equal justification, be called a compulsive personality or a paranoid one. Stagner and Moffitt (1956) have shown that persons high on a given personality trait do not show more

⁴ Gough (personal communication, 1960) found, in 100 military officers, that those high on the Strong vocational interest keys for "banker" and for "mortician" reported early ages of first sexual intercourse, and were described as coarse, noisy, and masculine. It is to be predicted that this group should score low on the factor of repression or hysterical character, and at the "harria" pole of Cattell's "premsia versus harria" factor. While the evidence of the present study is ambiguous, it is consistent with clinical impressions that hysterical character has a complex effect on sexual behavior, delaying age of first sex experience, at least in males, yet facilitating later unconsciously motivated promiscuity, "hysterical acting out" (Table D, in women).

than chance resemblance in other traits. These considerations make it doubtful that psychiatric diagnoses are comparable to medical ones. Choice of diagnostic label seems not to be a fruitful point of dispute.

In almost every instance, *K* corrected scales proved to be factorially purer than the same scales without *K*. This finding reinforces Fricke's (1956) recommendation that *K* corrected scores be used in factor analysis. The increased factorial purity of *K* corrected scales appeared only after rotation to simple structure, and was not found in the centroid factors before rotation. Since *K* corrected scores are more valid (Meehl & Hathaway, 1946), this may be evidence that rotation to oblique simple structure produces factors more genuine than the unrotated ones.

The extraction of factors that have a clear clinical and psychodynamic meaning is a satisfying result. It provides some rapprochement between psychoanalytic and experimental approaches. Even more satisfying is the fact that several of the factors correspond to defenses of the ego. For of all psychoanalytic concepts, defenses are most amenable to objective study—far more so than libidinal fixation, for example. Defenses have effects in overt behavior and in perception, effects that can be studied through experiments on observable behavior (and to some extent have been, e.g., Eriksen, 1952, 1954; Mowrer—in rats—1940). Measurements of defenses by questionnaire scales derived from factorial studies could be useful in such behavioral studies. The clinical usefulness of such measurements also appears promising.

SUMMARY

A Thurstone centroid factor analysis was done on 59 variables, including the original and some newer MMPI scales, on a group of 50 men, a group of 50 women, and the combined group of 100. Twelve or 13 factors were found in each group. The first 2 factors in all three groups resembled those previously found by Welsh and named *A* and *R*. These 2 factors were related to response sets, respectively, willingness to admit sickness and naysaying.

After rotation to oblique simple structure, an entirely different set of factors appeared. About six of these were replicated, being

virtually identical in the three samples. Four of these corresponded to ego defenses of reaction formation, repression, projection, and conversion. Alternatively, the first three could be named anal or compulsive character, hysterical character, and paranoid character. The highest loadings on these factors, among previously published scales, were, respectively, Gough's rigidity-flexibility, Wiener's *Hy*-subtle, and Wiener's *Pa*-subtle. New experimental scales had higher loadings on some factors.

The anal character factor also confirmed Freud's concept, having substantial loadings on three experimental scales for orderliness, stinginess, and stubbornness.

A fifth common factor, loaded with *Pd* and with Gough's delinquency-socialization, was tentatively identified with oral aggression. A sixth factor was marked by *Pt*, and named obsessive worrying, but in the combined sample bore more resemblance to a factor of general neuroticism.

REFERENCES

- ABRAHAM, K. A short study of the development of the libido, viewed in the light of mental disorders. (Originally printed 1924) In, *Selected papers on psychoanalysis*. New York: Basic Books, 1953. Ch. 26.
- ADORNO, T. W., FRENKEL-BRUNSWIK, E., LEVINSON, D. J., & SANFORD, R. N. *The authoritarian personality*. New York: Harper, 1950.
- AMERICAN PSYCHIATRIC ASSOCIATION, Committee on Nomenclature and Statistics. *Diagnostic and statistical manual of mental disorders*. Washington, D. C.: APA Mental Hospital Service, 1952.
- BARRON, F. X. An ego-strength scale which predicts response to psychotherapy. *J. consult. Psychol.*, 1953, 17, 327-333.
- CATTELL, R. B. *Factor analysis*. New York: Harper, 1952.
- CATTELL, R. B. *Handbook for the Sixteen Personality Factor Questionnaire*. Champaign, Ill.: Institute for Personality and Ability Testing, 1957. (a)
- CATTELL, R. B. *Personality and motivation structure and measurement*. Yonkers, N. Y.: World Book, 1957. (b)
- COOK, E. B., & WHERRY, R. J. A factor analysis of MMPI and aptitude test data. *J. appl. Psychol.*, 1950, 34, 260-265.
- COTTLE, W. C. A factorial study of the multiphasic, Strong, Kuder, and Bell inventories using a population of adult males. *Psychometrika*, 1950, 15, 25-47.
- COUCH, A. S., & KENISTON, K. Yeasayers and naysayers: Agreeing response set as a personality variable. *J. abnorm. soc. Psychol.*, 1960, 60, 151-174.

- DESOTO, C. B., & KUETHE, J. L. The set to claim undesirable symptoms in personality inventories. *J. consult. Psychol.*, 1959, 23, 496-500.
- DICKMAN, K. W. *Oblimax rotation of factors*. Urbana, Ill.: University of Illinois, Digital Computer Laboratory, 1959. (Mimeo)
- DRAKE, L. E. A social I.E. scale for the MMPI. *J. appl. Psychol.*, 1946, 30, 51-54.
- EDWARDS, A. L. *The social desirability variable in personality assessment and research*. New York: Dryden, 1957.
- ERIKSEN, C. W. Individual differences in defensive forgetting. *J. exp. Psychol.*, 1952, 44, 442-446.
- ERIKSEN, C. W. Some personality correlates of stimulus generalization under stress. *J. abnorm. soc. Psychol.*, 1954, 49, 561-565.
- FENICHEL, O. *The psychoanalytic theory of neurosis*. New York: Norton, 1945.
- FREUD, A. *The ego and the mechanism of defense*. New York: International Univer. Press, 1946.
- FREUD, S. *The problem of anxiety*. New York: Psychoanalytic Quarterly Press, 1936.
- FREUD, S. Character and anal erotism. (Originally printed 1908) In, *Collected Papers*. Vol. 2. London: Hogarth, 1950. Ch. 4.
- FRICKE, B. G. Conversion hysterics and the MMPI. *J. clin. Psychol.*, 1956, 12, 322-326.
- GOUGH, H. G. The *F* minus *K* dissimulation index for the MMPI. *J. consult. Psychol.*, 1950, 14, 408-413.
- GOUGH, H. G. *Manual for the California Psychological Inventory*. Palo Alto, Calif.: Consulting Psychologists Press, 1957.
- GOUGH, H. G., & PETERSON, D. R. The identification and measurement of predispositional factors in crime and delinquency. *J. consult. Psychol.*, 1952, 16, 207-212.
- HATHAWAY, S. R. *The Minnesota Multiphasic Personality Inventory manual*. (Rev. ed.). New York: Psychological Corporation, 1951.
- HECHT, S. An investigation into the psychology of masochism. Unpublished doctoral thesis, University of California, Berkeley, 1952.
- JONES, E. Anal-erotic character traits. In, *Papers on psychoanalysis*. New York: Wood, 1913.
- KARSON, S., & POOL, K. B. The construct validity of the Sixteen Personality Factors Test. *J. clin. Psychol.*, 1957, 13, 245-252.
- KENDALL, E. The validity of Taylor's manifest anxiety scale. *J. consult. Psychol.*, 1954, 18, 429-432.
- LAUTERBACH, C. G. The Taylor scale and clinical measures of anxiety. *J. consult. Psychol.*, 1958, 22, 314.
- LAZARUS, R. S., ERIKSEN, C. W., & FONDA, C. P. Personality dynamics and auditory perceptual recognition. *J. Pers.*, 1951, 19, 471-482.
- LINGOES, J. C. MMPI factors of the Harris and the Wiener subscales. *J. consult. Psychol.*, 1960, 24, 74-83.
- LITTLE, K., & FISHER, J. Two new experimental scales of the MMPI. *J. consult. Psychol.*, 1958, 22, 305-306.
- LUCHINS, A. R. Mechanization in problem solving: The effect of *Einstellung*. *Psychol. Monogr.*, 1942, 54(6, Whole No. 248).
- McKINLEY, J. C., & HATHAWAY, S. R. The Minnesota Multiphasic Personality Inventory: V. Hysteria, hypomania, and psychopathic deviate. *J. appl. Psychol.*, 1944, 28, 153-174.
- McNEMAR, Q. *Psychological statistics*. New York, Wiley, 1949.
- MEEHL, P. E., & HATHAWAY, S. R. The *K* factor as a suppressor variable in the Minnesota Multiphasic Personality Inventory. *J. appl. Psychol.*, 1946, 30, 525-564.
- MOWRER, O. H. An experimental analogue of "regression" with incidental observations on "reaction-formation." *J. abnorm. soc. Psychol.*, 1940, 35, 56-87. (Reprinted: In, *Learning theory and personality dynamics*. New York: Ronald, 1950. Ch. 13.)
- NAVRAN, L. A rationally derived MMPI scale to measure dependence. Unpublished doctoral dissertation, Stanford University, 1952.
- PINZKA, C., & SAUNDERS, D. R. Analytic rotation to simple structure: II. Extension to an oblique solution. *ETS res. Bull.*, 1954, No. 54-31.
- RIESMAN, D., GLAZER, N., & DENNEY, R. *The lonely crowd*. New Haven: Yale Univer. Press, 1950.
- ROKEACH, M. Generalized mental rigidity as a factor in ethnocentrism. *J. abnorm. soc. Psychol.*, 1948, 43, 259-278.
- ROSENZWEIG, S. The picture-association method and its application in a study of reactions to frustration. *J. Pers.*, 1945, 14, 3-23.
- SEARS, R. R. *Survey of objective studies of psychoanalytic concepts*. New York: Social Science Research Council, 1942.
- STAGNER, R., & MOFFITT, J. W. A statistical study of Freud's theory of personality types. *J. clin. Psychol.*, 1956, 12, 72-74.
- TAYLOR, J. A personality scale of manifest anxiety. *J. abnorm. soc. Psychol.*, 1953, 48, 285-290.
- WELSH, G. S. Factor Dimensions *A* and *R*. In G. S. Welsh & W. G. Dahlstrom (Eds.), *Basic readings on the MMPI in psychology and medicine*. Minneapolis: Univer. Minnesota Press, 1956.
- WHEELER, W. M., LITTLE, K. B., & LEHNER, G. F. J. The internal structure of the MMPI. *J. consult. Psychol.*, 1951, 15, 134-141.
- WIENER, D. N. Subtle and obvious keys for the Minnesota Multiphasic Personality Inventory. *J. consult. Psychol.*, 1948, 12, 164-170.
- WIGGINS, J. S. Interrelationships among MMPI measures of dissimulation under standard and social desirability instructions. *J. consult. Psychol.*, 1959, 23, 419-427.
- WIGGINS, J. S., & RUMRILL, C. Social desirability in the MMPI and Welsh's factor scales *A* and *R*. *J. consult. Psychol.*, 1959, 23, 100-106.
- WILLIAMS, H. L., & LAWRENCE, J. F. Comparison of the Rorschach and MMPI by means of factor analysis. *J. consult. Psychol.*, 1954, 18, 193-197.

(Received June 23, 1960)

PLAY THERAPY LIMITS AND THEORETICAL ORIENTATION

HAIM G. GINOTT¹ AND DELL LEBO²

Child Guidance Clinic, Jacksonville, Florida

Child therapists agree that reasonable and consistent limits are necessary in play therapy with disturbed children. However, there are differences of opinion about the specific limits required in therapy. Some therapists (see Dorfman, 1951, p. 262) only set limits on activities that interfere with their ability to remain accepting of the child. They may allow a child to take toys home, break equipment, urinate on the floor, leave the playroom at will, or terminate treatment. Other therapists (Ginott, 1959) set limits on such activities. They only allow children to express themselves symbolically through words, play, and nonverbal gestures.

Judging from the literature, psychoanalytic therapists are less preoccupied with the problem of limits than are nondirective therapists. Extensive discussion of limits can be found in many nondirective publications (Axline, 1947; Bixler, 1949; Dorfman, 1951; Moustakas, 1953); only passing references to this subject appear in psychoanalytic writings (Schiffer, 1952; Slavson, 1952). On this basis, one might assume that psychoanalytic therapists employ fewer limits than nondirective therapists.

The authors know of no published study of the use of limits by therapists of different schools. The purpose of the present investigation was to discover whether limit setting is related to theoretical orientation. The experimental hypotheses were: (a) Therapists of different schools will not differ in the number of limits that they employ in play therapy. (b) Therapists of different schools will not

differ in the kind of limits that they employ in play therapy.

METHOD

A questionnaire containing 54 discrete limits³ was sent to 425 child guidance clinics and other agencies that treat children. The respondents were asked to identify themselves as being primarily nondirective, psychoanalytical, or "other," and to indicate (Yes, No, or Sometimes) whether they used a particular limit with children who were neither psychotic nor organic and between the ages of 3-10 years.

Questionnaires were returned by 227 play therapists;⁴ of these 100 considered themselves to be psychoanalytic, 41 nondirective, and 86 of "other" schools. The number of Yes, No, and Sometimes responses to each of the 54 limits was totaled for each therapeutic school.

RESULTS

The mean number of limits used "ordinarily" and "sometimes" by therapists of the three schools are reported in Table 1. These means do not differ significantly; hence the results support the hypothesis that therapists of different orientations do not differ in the number of limits that they employ in play therapy.

³ Copies of the questionnaire may be obtained from either of the writers.

⁴ Most of the respondents were psychologists, a few were social workers and psychiatrists.

TABLE 1
MEAN NUMBER OF LIMITS "ORDINARILY" AND "SOMETIMES" USED BY PLAY THERAPISTS OF THREE SCHOOLS

Limit	Psychoanalytic	Nondirective	"Other"
Ordinarily	25	26	27
Sometimes	12	9	11

¹ Now at New York University.

² Our thanks are expressed to Georgia Dreger and Dorothy Cohen for their assistance in compiling and checking the large array of data, and to Arthur Orgel for helpful advice.

TABLE 2
SIGNIFICANTLY DIFFERENT PERCENTAGES OF LIMITS USED BY PLAY
THERAPISTS OF THREE SCHOOLS

Limit	Use	School			χ^2
		Psychoanalytic	Nondirective	"Other"	
Enter playroom	No	35	54	45	11.064
	Yes	26	15	28	
	Some	39	31	24	
Pour water	No	43	54	30	9.965
	Yes	24	24	33	
	Some	18	12	16	
Paint cheap toys	No	58	78	57	19.593
	Yes	14	15	23	
	Some	28	7	19	
Paint costly toys	No	20	27	15	9.821
	Yes	53	61	54	
	Some	23	10	23	
Bring drinks, food	No	56	51	44	18.448
	Yes	4	20	23	
	Some	40	24	31	
Light matches	No	30	42	26	10.162
	Yes	30	34	36	
	Some	28	22	38	
Read books	No	54	71	48	18.321
	Yes	6	10	17	
	Some	40	19	35	
Do school work	No	35	54	22	23.668
	Yes	25	24	33	
	Some	40	22	44	
Break costly toys	No	9	19	6	10.965
	Yes	69	71	74	
	Some	21	10	16	
Hit therapist mildly	No	30	39	24	9.795
	Yes	42	42	54	
	Some	28	19	22	
Tie therapist	No	21	37	35	24.763
	Yes	44	54	44	
	Some	34	7	19	
Shoot therapist	No	16	31	15	12.004
	Yes	69	62	67	
	Some	15	7	16	
Fondle therapist	No	6	2	6	13.401
	Yes	62	83	65	
	Some	29	12	27	
Urinate or defecate	No	3	10	4	14.287
	Yes	75	71	86	
	Some	22	19	7	

Note.—Since a small number of items were not answered some columns do not total 100%.

To test the second hypothesis regarding the kinds of limits employed by different therapeutic schools, all responses to each of the 54 items were summed according to therapeutic orientation. These were converted into percentages and distributed into 3×3 chi squares. Of the 54 items, 14 achieved significance at better than the .05 level of confidence.⁵ (On the basis of chance alone no more than 3 items might have achieved such significance.) The statistically significant items are shown in Table 2. These results do not support the second hypothesis that therapists of different orientations use the very same kind of limits in play therapy.

DISCUSSION

Within the confines of the present study it is clear that therapists of the three approaches employ a similar number of limits in their work with children. While there are differences in the kinds of limits used by the three approaches, a considerable body of prohibitions are employed by all.

Thus, in the area of physical aggression against the therapist, practitioners of all schools concur to the same degree in prohibiting a child from squirting water on the therapist, painting his clothing, throwing sand, or forcefully attacking him. They differ, however, in that nondirectivists were significantly more permissive in allowing a child to shoot darts at or to hit the therapist and significantly less permissive in allowing a child to tie the therapist.

In the area of physical aggression against equipment, practitioners of all schools concur to the same degree in prohibiting a child from spilling sand, painting walls and furniture, starting fires, breaking windows and inexpensive toys, and throwing objects around the room. They differed, however, in that nondirectivists were significantly more permissive in allowing a child to pour much water into the sand box, paint and break expensive toys. The "other" therapists were the least permissive in allowing the painting of inexpensive toys.

⁵ Three items whose statistical significance depended upon the Sometimes entries were omitted.

In the area of socially unacceptable behavior, practitioners of all schools similarly prohibited a child from smoking, using racial slurs, speaking or writing profanities in the playroom, making obscene objects, painting his face or clothing, undressing, and masturbating. They differed, however, in that the "other" therapists were significantly less permissive in allowing a child to urinate or defecate on the floor.

In the area of safety and health, practitioners of all schools similarly prohibited a child exploding a whole roll of caps, climbing on high window sills, drinking dirty water, or eating mud, chalk, or fingerprints. They differed, however, in that the nondirectivists were significantly more permissive in allowing a child to light matches in the playroom.

In the area of playroom routines, practitioners of all schools similarly prohibited a child from taking home toys or clay objects, turning off the lights, leaving or overstaying, bringing in a friend, and talking to passers-by. They differed, however, in that the nondirectivists were significantly more permissive in allowing a child to decide whether or not to enter the playroom, to read books, and to do his school work there. The psychoanalytic therapists were significantly more permissive than members of the other two approaches in allowing a child to bring drinks and food into the playroom.

In the area of physical affection, practitioners of all schools similarly prohibited a child from sitting on their laps, hugging, and kissing them. They differed, however, in that nondirectivists were significantly less permissive in allowing a child to fondle them.

While 14 limits were used differently by the therapists of the present sample, the fact remains that practitioners of all schools concurred in the use of a large number of limits.

SUMMARY

This study aimed to discover whether limit setting in play therapy is related to the therapist's theoretical orientation. Responses to a 54-item questionnaire on limits have been received from 100 psychoanalytic, 41 nondirective, and 86 "other" play therapists.

The results indicated that therapists of varied orientations did not differ in the number of limits used. While some significant differences were found in the kind of limits used, a considerable body of limits was employed by all.

REFERENCES

- AXLINE, VIRGINIA M. *Play therapy*. Boston: Houghton Mifflin, 1947.
- BIXLER, R. E. Limits are therapy. *J. consult. Psychol.*, 1949, 13, 1-11.
- DORFMAN, ELAINE. Play therapy. In C. R. Rogers,

Client-centered therapy. Boston: Houghton Mifflin, 1951. Pp. 235-277.

GINOTT, H. G. The theory and practice of "therapeutic intervention" in child treatment. *J. consult. Psychol.* 1959, 23, 160-166.

MOUSTAKAS, E. C. *Children in play therapy*. New York, McGraw-Hill, 1953.

SCHIEFFER, M. Permissiveness versus sanction in activity group therapy. *Int. J. group Psychother.* 1952, 2, 255-261.

SLAVSON, S. R. *Child psychotherapy*. New York: Columbia Univer. Press, 1952.

(Received July 8, 1960)

THE CONSTRUCT VALIDITY OF THE EDWARDS PPS HETEROSEXUALITY SCALE

E. JERRY PHARES AND CALVIN K. ADAMS

Kansas State University

The Edwards Personal Preference Schedule (PPS) (1954) appears to be increasingly used both in diagnosis and research. It purports to measure 15 personality needs growing out of H. A. Murray's work. The outstanding characteristics of the PPS appear to be a forced-choice item format, an attempt to control for social desirability in item choice, and 15 needs which are not obviously or necessarily connected with clinical pathology.

At the present time, however, many of the 15 scales of the PPS have not been systematically investigated. An exception would be the Achievement scale. Several studies have investigated the relationship between the PPS Achievement scale and the McClelland n Ach measure (Bendig, 1957; Himelstein, Eshenbach, & Carp, 1958; McClelland, 1958; Marlowe, 1959; Melikian, 1958). Still other studies have investigated the construct validity of some of the scales. Bernardin and Jessor (1957) utilized the Autonomy and Deference scales of the PPS as a measure of dependency and generally confirmed the construct validity of those two scales. Gisvold (1958) using an Asch group situation measure of conformity behavior found a significant negative relationship between it and the PPS Autonomy scale. No relationship between PPS Deference and conformity behavior was found, however. Zuckerman and Grosz (1958) used the Sway Test (a measure of suggestibility) and found low swayers scored significantly higher on the Autonomy scale than high swayers. The Deference and Succorance scales did not differentiate the groups. Zuckerman (1958) defining rebelliousness on the basis of sociometric scores could find only partial support for the hypothesis that rebellious subjects would score high on the Autonomy, Dominance, and Aggression scales of the PPS. In another study

of PPS n Ach, Worell (1960) demonstrated that subjects high on this scale showed significant superiority over low subjects in two verbal learning situations. In a factorial investigation of the entire PPS, Levonian, Comrey, Levy, and Proctor (1959) found a discrepancy between what the PPS is designed to measure and its actual factorial item content.

The present study is an attempt at the construct validation of the Heterosexuality scale of the PPS as regards males. This variable is described by Edwards (1954) as follows:

To go out with members of the opposite sex, to engage in social activities with the opposite sex, to be in love with members of the opposite sex, to kiss those of the opposite sex, to be regarded as physically attractive by those of the opposite sex, to participate in discussions about sex, to read books and plays involving sex, to listen to or tell jokes involving sex, to become sexually excited (p. 5).

Therefore, given this definition of heterosexuality how might males who score high on this scale behave differently from males who score low? Or, stated another way, will behavior in experimental situations developed in the light of the characteristics of such a heterosexuality construct relate to behavior on the Heterosexuality scale of the PPS? If so, the construct validity of the scale will have, in part at least, been supported.

Selection of subjects for this study was based upon the administration of a scale to 170 males in two general psychology classes at Kansas State University. The scale consisted of the 28 items of the PPS Heterosexuality scale plus 22 buffer items selected at random from other PPS scales.¹ From

¹ There is some evidence that removing items from the context of a standardized test may alter the nature of the items and responses to them (Edwards, Wright, & Lunneborg, 1959).

this procedure a high heterosexual group was drawn composed of 20 subjects with scores between 16 and 28. A low group contained 20 subjects with PPS scores between 1 and 8. The same two groups of subjects performed in both phases of the study to be reported below.

PHASE 1

Relationships between needs and esthetic preferences would appear logical ones given the pervasive role needs are assumed to play. Such a relationship between *n* Achievement and esthetic preferences has been investigated by Knapp (1958). For our purposes, if high PPS heterosexual subjects find sex and sex related activities more congenial and pleasurable than do low PPS subjects, it seems reasonable to conclude that their esthetic preferences for photographs might reflect this. Therefore, it was predicted that high PPS heterosexual subjects would place a higher esthetic value on photographs involving sexual elements in varying degrees than would low PPS heterosexual subjects.

Procedure

Sixty black and white photographs were selected principally from various art and photography magazines, but also from other sources. On a common sense basis, the photographs were classified as either sexual or nonsexual. The former consisted of such subjects as female nudes, female facial portraits, "cheesecake," boy and girl holding hands or kissing, etc. The nonsexual photographs ranged from skylines of New York to street beggars, animals, children, etc. Each photograph was then examined by three judges who independently determined its sexual or nonsexual character. Using unanimous agreement as a criterion eight photographs were eliminated as ambiguous. Next, 47 unselected males were drawn from general psychology classes at Kansas State University. In three group sessions of about 15 subjects each, the remaining 52 photographs were consecutively projected onto a screen. The subjects were asked to rate each photograph on a five-point scale along a dimension of esthetic value. This enabled a value to be assigned to each photograph obtained by summing the ratings of the subjects and dividing by 47.

The remainder of the pre-experimental work consisted of arranging the above photographs into seven groups of six photographs each. Each group of six contained four nonsexual and two sexual photographs, all of approximately equal value. This procedure eliminated 10 photographs whose values were either too high or too low to easily fit into any of the seven groupings. Following this, each group of six photographs was pasted onto 21" x 28" white cards in two rows of three each. The positions of

the sex photographs were randomly distributed throughout the series.

In the experiment itself, subjects in both groups were told they were participants in a study to develop an art appreciation test and were asked to rank the photographs in each group from one to six in terms of how artistically pleasing they were. Subjects were allowed a maximum of 14 minutes to examine each group of six, whereupon they made their rankings and the next group of six was presented. Each subject was given a score which consisted of the sum of his rankings of the 14 sex pictures.

Results

The mean and standard deviation of the high heterosexual group were 45.2 and 12.8, respectively, while the corresponding mean and sigma for the low group were 57.6 and 9.1.² This mean difference of 12.4 with a SE of the difference between means of 3.5 yields a *t* value of 3.5, significant at the .001 level. It is interesting to note also that only two cases in the high group exceeded the median of the low group and only one case exceeded the ninetieth percentile of the low group.

From these data it is clear that subjects high on the PPS Heterosexuality scale place a higher value on photographs involving sexual elements than do subjects low on that scale.

PHASE 2

In this study it was hypothesized that subjects high on the PPS Heterosexuality scale would show greater retention of material encouraging the importance of sexual information, dating, and sex education in the prevention of mental illness than would low PPS Heterosexuality subjects. This prediction was based on the assumption that subjects should better learn and retain material supportive of their needs than material opposed to them or that contravalant material should be more disruptive of the retention process than supportive material. This prediction would also be consistent with research on attitudes and their role in learning and retention (Edwards, 1941; Levine & Murphy, 1943).

Procedure

As a warmup, subjects were given a brief and easy digit span test. Subjects were then read two short passages. Following each, they were given 4 minutes to write down as much of it as they could remember.

² The higher the score, the lower the esthetic value.

The first or neutral passage was Memory Selection (A) from the Wechsler Memory Scale (1945). The second selection was the sexual passage which read:

Dr. Paul Rogers/ the eminent/ New/ London/ psychiatrist/ recently/ spoke/ on the importance of/ sex education/ and dating/ in the prevention/ of mental illness./ He urged/ males/ to learn/ much/ more about sex/ and sexual functions./ He also stressed/ the importance/ of active dating/ and experience with girls/ in promoting/ happiness./ The speech was/ very/ well received/ and increased/ his stature/ as an authority/ on sex./

Each subject's score was the total number of units recalled on the sexual passage subtracted from the total number recalled on the neutral passage. To eliminate negative numbers a constant of 10 was added to subjects' scores. This procedure enabled the control of such extraneous variables as intelligence and learning ability which might have accounted for differential retention in the two groups. Twelve protocols were randomly selected from the two groups of subjects to establish interjudge scoring reliability. For the Wechsler passage the interjudge scoring reliability coefficient for the two independent judges was .97, and for the sex passage, .92. The senior experimenter's scoring alone was used in testing the hypothesis.

Results

The mean retention score and standard deviation of the high heterosexual group were 10.95 and 3.8, respectively, and the corresponding mean and sigma for the low group were 12.0 and 2.8.³ The difference between the means is thus 1.05 and the *SE* of the difference between means 1.06. A *t* test gave a value of .9906. This result, while not significant, is in the predicted direction.

DISCUSSION

The hypothesis of the first study was confirmed in a highly significant manner. The second study produced nonsignificant results although in the expected direction. In view of the latter consideration, a *t* test was run again on the retention data after eliminating the five lowest scoring high heterosexual subjects and the five highest scoring low subjects. In this fashion, three subjects with PPS scores of 16, two with scores of 17, and five with scores of 8 were thrown out thus making more extreme groups of 15 each. The number of cases eliminated was determined arbitrarily

³ The higher the score, the lower the retention of sex material.

without reference to any distribution characteristics. Nor were there any "holes" likely to inflate a trend. As it happened, this procedure entirely eliminated all subjects with scores of either 16, 17, or 8—there was no need to "choose" which subjects to include and which to exclude. This procedure resulted in mean retention scores of 10.0 and 12.2 for the high and low groups, respectively. The *t* value, based on a mean difference of 2.2 and a *SE* of the difference between means of 1.15, was now 1.9, significant at the .06 level. Matching the groups pair-wise on the basis of their scores on the neutral passage and assuming the discrepancy scores to be correlated, lowered the obtained *p* level to only .05. These latter findings suggest either that the Edwards Heterosexuality scale is somewhat nondiscriminating except at the extremes or that our retention measure was not as sensitive as it might have been. In retrospect it seems that we could have constructed a more threatening or need-engaging sex paragraph or perhaps a measure more easily capable of being scored for distortion.

As a test of whether the two criterion measures were related within subjects, subjects' scores on the retention task were correlated with their scores on the photograph task. The correlation was only .155. Again however, eliminating the aforementioned 10 subjects raised the *r* to .29. This coefficient is not significant, but in view of the increase may suggest a modicum of generality, especially assuming a more sensitive retention measure.

All in all, the strong evidence provided by the results of the first hypothesis and the suggestive evidence of the second hypothesis appear to give support to the construct validity of the Edwards PPS Heterosexuality scale as regards males and underscore particularly its potential usefulness for research purposes.

SUMMARY

In this study the Edwards PPS Heterosexuality scale was investigated with respect to its construct validity. Groups of high and low scoring males were utilized. Two investigations, one involving esthetic preferences and the other retention scores, were carried out. In the first case it was found that high PPS males placed a significantly higher es-

thetic value on sexual photographs than did low subjects. In the second study suggestive evidence was found that high subjects exhibit better retention of sexual material than do low subjects.

Generally, the evidence appears supportive of the construct validity of the Edwards PPS Heterosexuality scale.

REFERENCES

- BENDIG, A. W. Manifest anxiety and projective and objective measures of need achievement. *J. consult. Psychol.*, 1957, 21, 354.
- BERNARDIN, A. C., & JESSOR, R. A construct validation of the Edwards Personal Preference Schedule with respect to dependency. *J. consult. Psychol.*, 1957, 21, 63-67.
- EDWARDS, A. L. Political frames of reference as a factor influencing recognition. *J. abnorm. soc. Psychol.*, 1941, 36, 34-50.
- EDWARDS, A. L. *Personal Preference Schedule*. New York: Psychological Corporation, 1954.
- EDWARDS, A. L., WRIGHT, C. E., & LUNNEBORG, C. E. A note on "social desirability as a variable in the Edwards Personal Preference Schedule." *J. consult. Psychol.*, 1959, 23, 558.
- GISVOLD, D. A validity study of the autonomy and deference subscales of the EPPS. *J. consult. Psychol.*, 1958, 22, 445-447.
- HIMELSTEIN, P., ESCHENBACH, A. E., & CARP, A. Interrelationships among three measures of need achievement. *J. consult. Psychol.*, 1958, 22, 451-452.
- KNAPP, R. H. n achievement and aesthetic preference. In J. W. Atkinson (Ed.), *Motives in fantasy, action, and society*. New York: Van Nostrand, 1958.
- LEVINE, J. M., & MURPHY, G. The learning and forgetting of controversial material. *J. abnorm. soc. Psychol.*, 1943, 38, 507-515.
- LEVONIAN, E., COMREY, A., LEVY, W., & PROCTOR, D. A statistical evaluation of Edwards Personal Preference Schedule. *J. appl. Psychol.*, 1959, 43, 355-359.
- MCCLELLAND, D. Methods of measuring human motivation. In J. W. Atkinson (Ed.), *Motives in fantasy, action, and society*. New York: Van Nostrand, 1958.
- MARLOWE, D. Relationships among direct and indirect measures of the achievement motive and overt behavior. *J. consult. Psychol.*, 1959, 23, 329-332.
- MELIKIAN, L. H. The relationship between Edwards' and McClelland's measures of achievement motivation. *J. consult. Psychol.*, 1958, 22, 296-298.
- WECHSLER, D. A standardized memory scale for clinical use. *J. Psychol.*, 1945, 19, 87-95.
- WORELL, L. EPPS n achievement and verbal paired associates learning. *J. abnorm. soc. Psychol.*, 1960, 60, 147-150.
- ZUCKERMAN, M. The validity of the Edwards Personal Preference Schedule in the measurement of dependency-rebelliousness. *J. clin. Psychol.*, 1958, 14, 379-382.
- ZUCKERMAN, M., & GROSZ, H. J. Suggestibility and dependency. *J. consult. Psychol.*, 1958, 22, 328.

(Received July 11, 1960)

SECOND-ORDER PERSONALITY FACTOR STRUCTURE IN THE OBJECTIVE TEST REALM¹

R. B. CATTELL

University of Illinois

R. R. KNAPP

United States Naval Personnel Research Activity, San Diego, California

AND I. H. SCHEIER

University of Illinois

A comprehensive, empirical system of personality factor dimensions based on factor analyses of a battery of objective tests has been described in previous publications (Cattell, 1955a; Cattell, 1957a). The initial emphasis in the development of this system has been to establish these as *first-order factors*, that is factors which are based on correlations between *tests*, and which therefore adhere closely to the detailed information available in tests. Recently, as first-order factor description has attained increasing replication, permitting more exact conceptualization and measurement, it has become possible to shift attention to *second-order factors*, that is factors based on the correlations between first-order factors (Cattell, 1956, 1957a). These correlations may be determined either (a) by measuring the factors by factor batteries, in which case they are attenuated by test unreliability and disturbed by invalidity, or (b) from the plots of rotations to simple structure in the first-order domain. In the latter case, there is no attenuation effect and relatively slight disturbance by invalidity, if and as the rotation is good and enough variables are employed to define the hyperplanes. Research needs to take both approaches, since the degree of consensus throws light on the magnitude of error. Both approaches are considered in this article.

Since factor analysis is a parsimonious operation, there will be fewer second-order

¹ The opinions expressed are those of the writers and are not necessarily shared by the Department of the Navy.

than first-order factors, just as there are fewer first-order factors than there are tests (Cattell, 1952). Second-order factors may therefore be looked upon as relatively broad descriptive categories, interpretable as representing general organizing influences in personality. Due to the limited number of explanatory concepts which the human mind seems able to juggle at one time, as well as to the analytic defects of behavioral analysis at the level of premetric general and clinical observation, the concepts of most psychologists fit factors at the second-order rather than the first-order level. For example, the personality concept measured by the questionnaire second-order extraversion factor in the 16 PF (Cattell, 1956; Cattell, Saunders, & Stice, 1957) is probably referred to more often than are concepts measured by the first-order components: Cyclothymia, Surgency, and Parmia (Cattell, 1957a; Cattell, Saunders, & Stice, 1957). Both levels have their use and, although secondaries (second-orders) lose some of the predictive power of primaries (first-orders), knowledge of the second-order structure is extremely important for understanding personality, developmentally and in action.

The position achieved by research to date, in respect to orders and media of observation, is as follows:

1. The establishment of first-order personality factors in the questionnaire (verbal, self-evaluative) and life record (behavior rated *in situ*) media of measurement (e.g.

Cattell, 1957a; Cattell, Saunders, & Stice, 1957).

2. The tentative establishment of first-order personality factors in the objective test (or T) realm (Cattell, 1957a; Cattell & Scheier, 1959; Scheier & Cattell, 1958)—where “objective” tests are understood as being tests which are relatively disguised in purpose, difficult to fake, and which are based on the subject’s performance in miniature situational tests rather than self-report (Cattell, 1958; Scheier, 1958).

3. The establishment of second-order personality factors in the questionnaire (or Q) medium of measurement (Cattell, 1956).

4. The establishment of *some* matches across Q and T media. After separate factorization within each of the two media (in 1 through 3 above), factor scores are correlated to seek matches in the two series. Results in this area are not yet definitive, but they strongly indicate that (a) at least four second-order factors in questionnaires match (i.e., measure the same dimensions as) first-order factors in objective tests (Cattell & Scheier, 1959, 1961; Scheier & Cattell, 1958), and (b) even when the entire second-order questionnaire factor realm is accounted for and matched as best one can with objective test factors, all but five or six of the objective test factors lack substantial questionnaire association—i.e., they do *not* have questionnaire factor equivalents at either the first- or second-order Q level. A possible inference is that many objective test factors are getting at areas of personality which questionnaire factors do not, and probably never can, measure.

5. First explorations have been made of the second-order factors among the objective test medium primaries. As Paragraph 4 above indicates, these might be expected to represent broader influences than those in questionnaire and rating primaries. The present paper concerns itself with organizing the evidence from these recent studies.²

² Eventually, second-order objective test factors will have to have their relations checked with first- and second-order Q factors. Presumably, if the evidence of Paragraph 4a above holds up, no direct matches will be found because second-order objective test factors occur at a higher order than any known

METHOD

The researches available for collation are five in number, four being based on method of analysis (b) and one on method (a) as described in the first paragraph of this paper. All studies operated upon individual difference patterns—i.e., they were not incremental or P technique studies concerned with state factors, but rather, R technique analyses concerned with trait factors (Cattell, 1952, 1957a). For close comparability, not confusing population differences with experimental error, the studies considered in the present paper were all based on young male American adults. For ease of reference in subsequent tables and discussion, they are referred to by temporary symbols used in other publications from the senior author’s laboratory (Cs, Co, etc., below). These studies are described in necessary detail elsewhere (Cattell, 1955b; Cattell & Scheier, 1959, 1961; Scheier & Cattell, 1958), but the essential characteristics are as follows:

Ca. 500 United States Air Force males measured on 128 variables. Rotation to oblique simple structure at the first-order yielded 16 factors (Cattell, 1955b) and was followed by a second-order analysis (Method b) of the correlations among first-orders (see Appendix 12 in Cattell, 1957a).

Ca. 250 United States Air Force males measured on 64 variables. Rotation to oblique simple structure at the first-order yielded 15 factors (Cattell, 1955b) and was followed by a second-order analysis (Method b) of the correlations among first-orders (see Appendix 12 in Cattell, 1957a).

R₁, R₂. Two studies each of which measured the same 86 male college undergraduates, with 120 and 103 variables, respectively (69 of which were common to both studies). Rotation to oblique first-order simple structure, reported on elsewhere (Cattell & Scheier, 1959; Scheier & Cattell, 1958), yielded, respectively, 15 and 17 factors, and was followed by second-order analysis (Method b) of the correlations among first-order factors. These two second-order resolutions are published here for the first time.

N₁. 315 United States Navy male Submarine School candidates³ were scored on 18 first-order objective test factors as measured by the O-A Battery (Cattell, 1955a). That is, in the N₁ study each person’s score on each factor was obtained from a combination of test scores demonstrated, by consistent and substantial loadings on each factor, to give a best possible estimate of that factor. Correlations were computed among these battery scores (Method a above), and factors were then extracted and rotated to oblique simple structure. The second-order results are reported for the first time. The N₁ study is thus unique in that *factor score* correlations were used instead of the C_F matrix (Cattell, 1952) derived from

Q factors, namely at a *third-order*, relative to first-order Q factors.

³ These data were collected at the United States Naval Medical Research Laboratory, New London, Connecticut, under Bureau of Medicine and Surgery Project NM 23 02 20.

oblique rotations, as was the case in the other four studies.

Second-order factorization in all of the five studies used tests of completeness of factor extraction described elsewhere (Cattell, 1952) and pursued simple structure blindly to a plateau at which hyperplane count proved unimprovable. At that point the factor patterns were inspected and a process of matching (correlation of loading patterns) was begun, factor by factor, with each of the other four studies. This matching revealed tolerably good convergence of results in the five studies and the matched factors and their loadings are therefore set out in Tables 1 through 7 below, each table representing a second-order objective test factor which we now believe to be reasonably well confirmed. All five studies replicate the same seven second-order factors, and although a given factor is often less clear in one series than in another, no alternative matching would be nearly as satisfactory as that presented.

The layout of these tables follows the usual conventions, as follows:

1. Each table is headed by a contingent verbal title for the second-order factor and a roman numeral. The verbal title may be modified as more information becomes available on the factor, but its roman numeral remains constant for easy identification in present and future studies. Each verbal title is bipolar, the positive (high score) pole being on top and the negative in parentheses.
2. First-order factors—the "variables" at this order of analysis—are identified in the left-hand column of each table, (a) by their Universal Index or UI numbers (Cattell, 1957b), which are constant across all published researches, and (b) by their verbal titles.
3. The figures in Columns 3 through 7 are the loadings of first-orders on the second-order, each column being designated by the symbol for the reverse in which the results were found. These values differ from the true values, in the case of C_6 , C_6 , R_1 , and R_2 , by reason of imperfect simple structure in first- and second-order relations, and in the case of N_1 by attenuation due to unreliability of the factor batteries. Inconsistencies of sign between

one study and others are indicated by enclosing the atypical value in brackets. A positive sign on a loading means that the positive pole of that factor, as it is standardly written in various textbooks (Cattell, 1957a; Cattell & Scheier, 1961) goes positively with the second-order factor pole as written, and oppositely for a minus loading. Since, in reading titles on the left, readers like to see at once which "go together," we have followed the practice of writing in the pole which is consistent with the positive direction of the second-order factor. That is to say, when a loading in the numerical columns is negative (for the positive pole of the first-order factor), we have already reversed the title to put the negative pole of the factor in this title column. Therefore, the verbal title of the first-order already gives the direction in which (pole at which) it is related to the second-order factor. The signs on the loadings merely tell whether this is the positive (high score) or negative (low score) pole of the first-order, as it is usually scored and thought of.

4. The second column gives the average loading, which, by reason of the attenuation in N_1 , would be expected to err systematically, slightly below the true value. The given rank order in listing the primaries is not literally the declining order of their mean loading on the second-order, but an estimate of the order of importance of the primaries, made by taking into account also the consistency and frequency of replication of the result.

5. The tables report data only on the first-orders which are most highly and consistently associated with second-orders. The full tables, showing the loading of all first-orders present in each study, including those with lesser or essentially zero relationship, are preserved in tables available from the American Documentation Institute.⁴

⁴ Complete loading tables for all first-order factors have been deposited with the American Documentation Institute. Order Document No. 6758 from the ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance \$1.25 for microfilm or \$1.25 for photocopies. Make checks payable to: Chief, Photoduplication Service, Library of Congress.

TABLE 1

F(T)I: TIED SOCIALIZATION OR SUPREGO (VS. ABSENCE OF CULTURAL INTROJECTION)

First-Order Factor	Loadings in Studies					
	Average	C ₆	C ₆	R ₁	R ₂	N ₁
UI 20 Comention, Conformity to Cultural Standards	+36	+62	+46	+03		+34
UI 28 Rigid Superego	+26	+64	+13		+22	+05
UI 1 Low General Intelligence	-34	-34	-33			
UI 19 Subduedness	-20	-30	-37	-09	-06	-20
UI 25 Accurate Realism	-33	-04		-52	-15	-62
UI 35 Self-Reliance	+44			+44		
UI 32 Extraversion	-19	-17	-38	-01	-21	-16

TABLE 2

F(T)II: EXPANSIVE EGO (VS. HISTORY OF DIFFICULTY IN PROBLEM SOLVING)

First-Order Factor	Loadings in Studies					
	Average	C _s	C ₆	R ₁	R ₂	N ₁
UI 16 Harric Assertiveness	+34	+47	+56	+32	+36	00
UI 1 High General Intelligence	+28	+42	+13			
UI 19 Promethean Will	+23	+18	+33	+21	+06	+37
UI 23 Ergic Regression	-21			-08	-06	-48
UI 36 Self-Sentiment Development	+29			+29		
UI 18 Naivete	-15	00	-25	-03	-26	-22

DATA AND INTERPRETATION FOR SECOND-ORDER FACTORS

Factor F(T)I, Tied Socialization, has claims to being the largest second-order factor. There is high internal psychological consistency (with the possible exception of low intelligence) in terms of the characteristics of the first-order factors involved. A clinician might well claim it to be proof of a broad superego organization, and some solid support for this lies in the finding (Cattell & Scheier, 1961) that neurotics are significantly higher than normals on this second-order factor. However, in the label "Tied Socialization" we have favored an hypothesis which *may* specialize to a superego definition, but which is at present broader. It implies that this pattern represents the extent to which the individual has accepted the culture patterns and standards of the group. It has overtones of extraversion and realistic contact, subduedness, receptivity, and lower intelligence (un-

critical?), but the central factors are UI 20, Comention (conventional acceptance of values) and UI 28, Rigidity of Superego, which we believe the contingent title aptly designates.

Except for UI 23(-), there is again considerable psychological consistency in F(T)II, here in the sense of willfulness ("will power") and a high development of the self-sentiment. Once again, too, the clinician might claim that this particular clustering of first-order factors gives support to an observational concept at the second-order level, in this case, the concept of a broad dynamic organization that is called "ego development" or "ego strength." The essential unifying concept, covering higher intelligence and the assertive characteristics, appears to be, if an environmental explanation is adopted, a history of success in managing ergic (drive) satisfactions. This interpretation would also fit psychoanalytic concepts of ego strength. The

TABLE 3

F(T)III: TEMPERAMENTAL ARDOR (VS. TEMPERAMENTAL APATHY)

First-Order Factor	Loadings in Studies					
	Average	C _s	C ₆	R ₁	R ₂	N ₁
UI 21 Exuberance, Energetic Spontaneity	+31	+70	+40	+23	+11	+13
UI 20 Comention, Conformity to Cultural Standards	+27	+48	+12	+08		+38
UI 1 Low General Intelligence	-28	-15	-41			
UI 19 Promethean Will	+21	+21	+12	+18	+54	+01
UI 27 Alert Control	-19				-29	-09
UI 32 Extraversion	-16	(10)	-34	-55	(03)	-06

TABLE 4

F(T)IV: HIGH EDUCATED SELF-CONSCIOUSNESS (VS. LOW EDUCATED SELF-CONSCIOUSNESS)

First-Order Factor	Loadings in Studies					
	Average	C _s	C _e	R ₁	R ₂	N ₁
UI 22 Corticalertia, Cortical Alertness	+31	+64	+40	+08	(-02)	+47
UI 18 Shrewdness	+28	+66	+24	+40	+07	+01
UI 25 Imaginative Tension	+17	+45		+12	+02	+08
UI 36 Self-Sentiment Development	+51			+51		
UI 30 High General Reactivity	-16	-03	-08		(+05)	-57
UI 29 Low Adaptation Energy	-15	-01	-08		-53	(+04)
UI 33 Dourness	+20					+20

chief criterion evidence on this factor shows it distinguishing even more powerfully than F(T)I between neurotics and normals, the normals being significantly higher than neurotics in Expansive Ego (Cattell & Scheier, 1961).

A noteworthy feature of F(T)III is that it tends to pick out those factors which individually appear to have relatively high genetic determination (Cattell, 1957a). Previous multivariate experiments (Cattell, Stice, & Kristy, 1957) using groups of identical and fraternal twins and siblings, some of which were reared together and some apart, have attempted to assess the contribution which heredity and environment may make to certain of the objective test factors. In particular it appeared that heredity was the main determiner in F(T)III Factors UI 1, Intelligence, and UI 20, Comention. Hereditary and environmental influences appeared to be about equal in determining Factors UI 19, Promethean Will, and UI 27, Alert Control, which are found on F(T)III. Conceivably, these genetic determinations all arise from a single

genetic influence covering all primaries contributing to this second-order factor. However, it is difficult to imagine in polygenic determination how a series of genes would happen to coincide in repeatedly affecting several things in the same way. This is a provocative finding for psychological genetics. Meanwhile, we shall label the second-order "Temperamental Ardor," since, except for UI 20, the central character is a willful exuberance and ardor of temper. Consistent with this interpretation, F(T)III has been found to be significantly higher for hospitalized neurotics of sociopathic type than it is for normals (Cattell & Scheier, 1961).

F(T)IV tends to be loaded by what appear to be largely environmentally determined factors. In previous investigations (Cattell, Stice, & Kristy, 1957) environment appeared to be the main determiner in F(T)IV Factors UI 22, Corticalertia, and UI 29, Low Adaptation Energy-vs.-Overresponsiveness. It probably has something to do with education in the sense of developing alert, shrewd, and imaginative qualities. It also involves much ex-

TABLE 5

F(T)V: HISTORY OF INHIBITING, RESTRAINING ENVIRONMENT (VS. LAXNESS)

First-Order Factor	Loadings in Studies					
	Average	C _s	C _e	R ₁	R ₂	N ₁
UI 17 Inhibition	+36	+32	+54	+45	+03	+47
UI 23 High Mobilization of Resources	+18			+25	+19	+10
UI 31 Wary Realism	+16	+07	(-12)			+52

TABLE 6

F(T)VI: NARCISTIC DEVELOPMENT (VS. ENVIRONMENTAL CONTACT AND INVESTMENT)

First-Order Factor	Average	Loadings in Studies				
		C ₅	C ₆	R ₁	R ₂	N ₁
UI 26 Narcistic Self-Will	+33	+59	+31	+50	+25	00
UI 27 Apathy-Fatigue	+30				+36	+23
UI 34 Autistic Nonconformity	+51			+51		

plicit self-awareness and "canniness." Possibly, an upbringing in a critical, competitive, high-standard home atmosphere could generate such a pattern. The above interpretation is, at least, not inconsistent with the fact that neurotics with marked sociopathic trends have significantly lower scores than normals on this factor (Cattell & Scheier, 1961).

Our hypothesis is that F(T)V represents a dimension of personality resulting from an environment in which considerable inhibition prevails. The somewhat unexpected role of UI 23, here as in F(T)II, suggests that there may be some need to modify our hypotheses about this primary. However, it is reasonable to interpret UI 23 as unused reserves of energy and it might then fit here as the cumulative result of inhibited, undischarged reactivity.

F(T)VI is a rather narrow factor which nevertheless has a consistent character of narcissistic and autistic development. It is in some sense a false ego development, a moving out of contact with reality. This hypothesis is supported by its being significantly high in neurotics with sociopathic trends, but not in typical neurotics (Cattell & Scheier, 1961).

In F(T)VII, we may have a form of tension wider than Free Anxiety, UI 24, and therefore perhaps representing total drive tension as it is oriented and controlled toward achievement. On the other hand, an argument could be made from present inclusions and exclusions that the central influence in the factor is insecurities connected with the self-concept. We shall keep interpretive eventualities open by a descriptive title indicating anxiety and related drive tensions centered on insecurity, and under control of an achievement goal.

Now that the patterns of second-order objective test factors have been presented and tentatively interpreted, it is logical to ask if very broad *third-order* factors can be found. Accordingly, we computed the correlations among second-order T factors, in the three most recent studies (R₁, R₂, N₁) and averaged these values in Table 8. These correlations among second-orders are generally very low. Over half of the *r*'s are .10 or less, and only one, between F(T)III and F(T)VI, could confidently be called statistically significant. It is possible that a third-order might be found, including the second-orders of Tem-

TABLE 7

F(T)VII: HIGH TENSION TO ACHIEVE, CONTROLLED DRIVE TENSION LEVEL (VS. LOW TENSION TO ACHIEVE)

First-Order Factor	Average	Loadings in Studies				
		C ₅	C ₆	R ₁	R ₂	N ₁
UI 24 Free Anxiety	+40	+38	+38	+52	+42	+30
UI 18 Shrewdness	+23	+14	+38	+34	+18	+09
UI 30 High General Reactivity	-19	-56	-12		00	-07
UI 25 Imaginative Tension	+18	+49		+14	(-07)	+17
UI 19 Promethean Will	+16	+13	+10	+07	+05	+46

TABLE 8
CORRELATIONS AMONG SECOND ORDER
OBJECTIVE TEST FACTORS

I	II	III	IV	V	VI	VII
	-.10	-.06	-.17	.19	-.18	.07
		.01	-.07	-.17	.07	.12
			-.18	-.14	.38	-.14
				-.08	-.06	.05
					-.08	-.07
						-.04

Note.—Average over R_1 , R_2 , and N_1 studies.

peramental Ardor—F(T)III—and Narcistic Development—F(T)VI; but such factorization should await further replication and confirmation of second-order results. However, even now, it is possible to make a general prediction that, for practical purposes, oblique factorizations do not permit an indefinitely large series of higher-and-higher order factorizations. The correlations definitely become lower as one moves to higher orders; that is, correlations among second-orders are generally lower than correlations among first-orders, and correlations among first-orders are generally lower than correlations between tests.

There is no logical reason why the same basic data should not be describable by using sets of categories, alternatively at lower or higher orders, with the broader categories being fewer but missing more detail. On the other hand, it would be somewhat confusing and discomfiting if this process were to prove empirically and practically possible over, say, five, six, or even a greater number of distinct levels of generality (orders). The data of Table 8 strongly suggest an upper limit of three orders for objective test data, and, since there are only four well-confirmed second-order factors in questionnaires (Cattell, 1956; Cattell & Scheier, 1961), it is almost certain that they too will yield no more than three orders.

SUMMARY

1. Correlations among 20 first-order personality factors, in objective tests, have been obtained in five separate studies employing young male adult samples.

2. An oblique simple structure factoring, independently on each of the five correlation matrices, yields seven factors, the general form of which is confirmed by matching across the five studies.

3. With relatively short supportive discussion, these factors have been indexed and named as follows:

- F(T)I, Tied Socialization
- F(T)II, Expansive Ego
- F(T)III, Temperamental Ardor
- F(T)IV, Educated Self-Consciousness
- F(T)V, History of Inhibiting, Restraining Environment
- F(T)VI, Narcistic Development
- F(T)VII, Tension to Achieve

4. Space dictates restricting this presentation to evidence of the patterns, their matching, and a few criterion associations. Elsewhere, a continuation will be made of the fuller theoretical development required by this first demonstration of consistent personality structure at the more pervasive level of second-order objective test factors.

5. Possibility of third-order factors is briefly discussed.

REFERENCES

- CATTELL, R. B. *Factor analysis: An introduction and manual for the psychologist and social scientist*. New York: Harper, 1952.
- CATTELL, R. B. *Handbook for the objective-analytic personality test batteries*. Champaign, Ill.: Institute of Personality and Ability Testing, 1955. (a)
- CATTELL, R. B. Psychiatric screening of flying personnel: Personality structure in objective tests—a study of 1,000 Air Force students in basic pilot training. *USAF Sch. Aviat. Med. proj. Rep.*, 1955, No. 9. (Project No. 21-0202-007) (b)
- CATTELL, R. B. Second-order personality factors in the questionnaire realm. *J. consult. Psychol.*, 1956, 20, 411-418.
- CATTELL, R. B. *Personality and motivation structure and measurement*. New York: World Book, 1957. (a)
- CATTELL, R. B. A universal index for psychological factors. *Psychologia*, 1957, 1, 74-85. (b)
- CATTELL, R. B. What is "objective" in "objective personality tests"? *J. counsel. Psychol.*, 1958, 5, 285-289.
- CATTELL, R. B., SAUNDERS, D. R., & STICE, G. *Handbook for the Sixteen Personality Factor Questionnaire*. Champaign, Ill.: Institute of Personality and Ability Testing, 1957.

- CATTELL, R. B., & SCHEIER, I. H. Extension of meaning of objective test personality factors: Especially into anxiety, neuroticism, questionnaire, and physical factors. *J. gen. Psychol.*, 1959, **61**, 287-315.
- CATTELL, R. B., & SCHEIER, I. H. *The meaning and measurement of neuroticism and anxiety*. New York: Ronald Press, 1961.
- CATTELL, R. B., STICE, G. F., & KRISTY, N. F. A first approximation to nature-nurture ratios for eleven primary personality factors in objective tests. *J. abnorm. soc. Psychol.*, 1957, **54**, 143-159.
- SCHEIER, I. H. What is an "objective" test? *Psychol. Rep.*, 1958, **4**, 147-157.
- SCHEIER, I. H., & CATTELL, R. B. Confirmation of objective test factors and assessment of their relation to questionnaire factors: A factor analysis of 113 rating, questionnaire and objective test measurements of personality. *J. ment. Sci.*, 1958, **104**, 608-624.

(Received July 13, 1960)

LEVEL OF ASPIRATION IN HYPERTENSIVE CARDIAC PATIENTS COMPARED WITH NONHYPERTENSIVE CARDIAC PATIENTS WITH ARTERIOSCLEROTIC HEART DISEASE

DESMOND D. O'CONNELL

Veterans Administration Hospital, Wood, Wisconsin

AND RICHARD M. LUNDY

University of Wisconsin

The present study¹ is concerned with the application of the "level of aspiration" technique to the study of persons with cardiac disease. Since 1930, when Hoppe first described the level of aspiration phenomenon, only six studies cited in *Psychological Abstracts* relate to physical or physiological conditions. Two studies—one reported by Berkeley (1952) and one reported by Gerard and Phillips (1953)—found a reliable relationship between adrenal activity and level of aspiration scores. Little and Cohen (1951) found that asthmatic children showed significantly higher levels of aspiration than non-asthmatic. Hecht (1952) was able to differentiate ulcer patients from colitis patients on level of aspiration scores, the ulcer patients getting significantly higher "D scores." Scodell (1953) also found differences between peptic ulcer patients and a neurotic (non-ulcer) control group on a level of aspiration test as well as on several other measures; the ulcer group had lower level of aspiration scores than the neurotic group. Raifman (1957) found that a group of ulcer patients showed significantly higher levels of aspiration than a group of normals and a group of neurotics. The present study will attempt to determine whether persons with hypertensive heart disease can be differentiated from

persons with nonhypertensive arteriosclerotic heart disease on a level of aspiration task.

Perhaps the best known effort to describe distinctive personality characteristics of persons with cardiac disease is that of Dunbar (1948) who described personality profiles for several types of patient. Alexander (1950) considers that probably the most valid of Dunbar's profiles is that of the "coronary"² patient. Yet an objective study of 46 coronary patients performed more recently by Miles, Waldfogel, Barrabee, and Cobb (1954), using psychiatric interview, social history, and psychological tests, disagrees with or fails to confirm most of the specific characteristics attributed to coronary patients by Dunbar. The study did show, however, more strenuous work histories, with more physical and psychological stress and strain than the group of normals; in this respect the study tends to confirm that part of Dunbar's profile which pictures the coronary patient as a consistently striving person who works harder and longer

² It will be noted that the word "coronary" in the introductory section has been used in quotation marks. This is because the term needs clarification; "coronary" and "hypertensive" are not mutually exclusive types of heart disease. Dunbar noted, for example, that 27% of the patients with coronary occlusion in her study also had hypertension (Dunbar, 1948, p. 251). Many hypertensives may eventually have coronary occlusion or coronary insufficiency. Miles, Waldfogel, Barrabee, and Cobb (1954) in their study of coronaries, limited the study to cases in which there had actually been a coronary attack—occlusion of the coronary artery—with an absence of hypertension.

¹ The authors wish to express their appreciation to Jules Chase, Consultant in Cardiology at the Veterans Administration Hospital, Wood, Wisconsin, for a suggestion made in the original design of the study and for his critical reading of the manuscript from the medical viewpoint.

than the average adult (Dunbar, 1948, p. 307). Alexander (1950, p. 72) agrees with Dunbar on this characteristic.

The person with hypertensive cardiac disease is characterized by Dunbar as having in common with coronary patients a constant striving to subdue or surpass competitors, but as being different from coronary patients in that hypertensives have a greater fear of criticism, a greater fear of responsibility, a greater fear of falling short. They are more likely to choose occupations below their ability and are usually less successful than coronaries (Dunbar, 1948, p. 264).

If hypertension is preceded by a history of being unsuccessful, fearful of criticism, or of responsibility, and of falling short, as suggested by Dunbar, it may produce a different pattern of response on a level of aspiration test than would be produced by a more successful, less fearful history. There is some support for this hypothesis in the literature. Sears (1940), for example, has shown that chronically unsuccessful children, as judged by school achievement, show a different pattern of response from successful children on a level of aspiration test. The pattern was not necessarily one of lowered aspirations, however, as might be suggested by Dunbar's assertion that the hypertensive is more likely to choose occupations below his ability. In Sears' study the unsuccessful group produced a bimodal distribution, showing either an unrealistically higher level of aspiration or an unnecessarily low level of aspiration. The successful individual typically sets his goal near but slightly above his last previous performance. Kurt Lewin (Lewin, Dembo, Festinger, & Sears, 1944) in his review of the literature and again in 1948 affirms these as characteristic reactions of the successful and the unsuccessful individual in level of aspiration situations.

In a recent theoretical article concerned with level of aspiration and risk taking behavior, Atkinson (1957) has drawn a distinction between those responses, or persons, whose motivation is to achieve success, and those whose motivation is to avoid failure. Those who seek to achieve success typically chose a task near their achievement level, i.e., where the uncertainty of success or failure

is greatest. Those who seek to avoid failure typically set their goals either very high or very low. Thus in a level of aspiration study we might be concerned with two kinds of responses: the achievement oriented and the failure oriented. These types of responses seem comparable to Scodel's "typical" and "atypical" responses. Similarly, Scodel (1953) divides the atypical responses into atypical high and atypical low responses.

If Dunbar is correct in her assertions about persons with coronary heart disease and those with hypertensive heart disease, the two groups should be similar when achievement oriented behavior is considered. The responses of the two groups in their failure oriented behavior, however, should differ, according to her view that the hypertensives are more fearful of failure. She also states that the direction of this failure oriented reaction is toward choosing atypically low levels of aspiration—in Atkinson and Scodel's terms.

Although the writings of Dunbar have given impetus to this general investigation, the findings of Miles et al. (1954), have cast considerable doubt about her specific hypotheses. The approach of the present investigation, however, is not to test Dunbar's specific hypotheses. Rather, we feel that the status of the theoretical positions is such that only an empirical approach is presently warranted. Thus, the questions asked in the present study are: do hypertensive heart cases differ from nonhypertensive, arteriosclerotic heart cases in (a) the number of achievement oriented, aspiration responses, and (b) the extent and direction of the failure oriented responses?

METHOD

Subjects

All subjects were adult male veterans between the ages of 32 and 65, hospitalized at Veterans Administration Hospital, Wood, Wisconsin. Patients in this hospital are from Wisconsin, Michigan, and Illinois; the largest percentage of patients are from the Milwaukee metropolitan area.

All new admissions to the hospital were reviewed and a list made of all those showing an admission diagnosis of heart disease or suspected heart disease. After the medical diagnostic procedures had been completed, the cases were discussed with the ward physician. Subjects with an established diagnosis of

"Hypertensive Heart Disease" were assigned to one group; subjects with an established diagnosis of "Arteriosclerotic Heart Disease" and an absence of hypertension were assigned to the second group. Subjects with other types of heart disease were eliminated from further consideration as not meeting the criteria of the defined groups. Among those meeting the definitional or categorical requirements no eliminations were made except subjects who were mentally or physically incapable of cooperating in the study (for example, subjects who have had a cardiovascular accident resulting in hemiplegia, or aphasia, or deterioration of mental functioning).

Medically there is always the possibility that a subject in the nonhypertensive arteriosclerotic group has at some time in the past been hypertensive. However, from the standpoint of research design, such a misclassification would not invalidate a positive finding. Thus, if there is a psychological variable associated with the presence of hypertension, and the present study indicates that the nonhypertensive group and the hypertensive group differ significantly on this variable, the significance level would be a minimal value; transfer of the misclassified case to the hypertensive group in which it belongs would only serve to increase the significance level.

Associated diseases found to be present coincident with the primary diagnoses of hypertensive or arteriosclerotic heart disease were of interest, but were not used in selecting the subjects because to do so would change the basis of selection and disturb the research design of the study.

Each subject was given 11 15-second trials on the Placing test of the Minnesota Rate of Manipulation Test. He was asked to set a goal prior to each trial. The test was presented to him as a test of his coordination.

This task was selected after consideration of a number of factors. The literature on the use of the level of aspiration technique reveals that a wide variety of tasks, both verbal and motor, have been used; tasks used, for example, include card sorting, cancellation, addition, pegboards, bowling games, target tests, and the Rotter Aspiration Board. A decision was made in favor of the Minnesota Placing test for the following principal reasons: (a) Hecht (1952) found significant differences between peptic ulcer and ulcerative colitis patients using the Purdue Pegboard; the Minnesota Placing test is a somewhat similar performance-type test of coordination but requires less finger dexterity and less precise placement. (b) Bowling games and other "game"-type tasks in level of aspiration studies have been criticized by Barnett, Handelsman, Stewart, and Super (1952), Stubbins (1950), and others as having too much of a "play" atmosphere and not being closely related to life situations. This has led to the practice in recent level of aspiration studies of introducing penalties or other extraneous motivation. If this criticism has any validity, it might be that a game-type task would be even less meaningful and challenging to the older individuals with which the present study deals. On the other hand, it was our feeling that the Minnesota Placing test would prove to be a meaning-

ful task when presented as "a test of coordination" to a hospitalized cardiac patient who has just come through a cautiously graded progression in amount of physical activity permitted and who is very much aware of and concerned about his physical limitations.

Test Instructions

Standardization of instructions was considered especially important in this study because it is known from previous studies of level of aspiration technique, such as those of Irwin (1942), of Walder (1951), and of Saji (1951) that the form of the stimulus question affects the results. Thus, "What score do you expect to get?" "What do you hope to get?" or "What will you try to get?" may bring different responses. Using Frank's (1941) definition of level of aspiration as, "the level of future performance in a familiar situation which an individual explicitly undertakes to reach," we have used as our stimulus question: "How many blocks will you try to get in on the next trial?" After the initial trial, the stimulus question was shortened to: "How many will you try for this time?"

Age

Age has not been shown to be a significant factor in level of aspiration studies; studies of this factor which have been done, such as those by Adams (1939), Walter (1948; Walter & Marzolf, 1951), and Reissman (1953), have tended to show that level of aspiration behavior is not significantly related to age. However, some of the studies did not examine age as a continuous variable but merely compared "old" groups with "young" groups; other studies which did examine age as a continuous variable have usually considered only a very limited range of ages. Because there has been insufficient research on age as a variable in level of aspiration behavior, and because age is known medically to be related to arteriosclerotic heart disease, it was our feeling that it should be controlled in the present study. The age of each subject was therefore recorded for comparison (see Results section).

Stage of Recovery

Although the effect of stage of recovery on level of aspiration behavior is unknown and uninvestigated, it seemed desirable to control this factor. The experimental task was light sedentary activity involving only 15 seconds of activity at a time and an over-all time of only about 15 minutes; yet it required rapid movement of the arm, and might be either a real or a perceived threat in the early stages of recovery from cardiac disease. Thus no patients were scheduled until they reached the "ambulatory" stage of recovery, and this was approximately the same for all subjects.

Education

The average education for the Hypertensive group is 8.7 with a standard deviation of 2.0, and a range of 5-13 years of formal schooling completed. This

TABLE 1

DIFFERENCE BETWEEN THE MEANS OF THE HYPERTENSIVE AND ARTERIOSCLEROTIC GROUPS USING THE "A" SCORES

Group	N	M	SD	SE _D	t
Hypertensive	24	5.75	2.57	0.72	1.83*
Arteriosclerotic	23	4.43	2.34		

* $p < .10$.

compares with an average of 8.8 and a standard deviation of 2.6 for the Arteriosclerotic group, with a range of 4-14 years completed.

Method of Scoring the Aspiration Responses

Although many systems of scoring have been used in various studies in the literature, the most widely used is the D score or the average discrepancy between the subject's performances and his aspirations. The D score, however, fails to take account of what would appear to be an important psychological variable: the effect of success or failure upon the subject's immediately subsequent aspiration. Thus, an average D score of +1 for one person may represent a very consistent aspiration to do one better than the last performance no matter whether it was a success or failure experience. For another person a +1 score could represent an average of quite different reactions following success or failure.

Guided by the Atkinson (1957) formulation that the achievement oriented individual sets his goal at the point of greatest uncertainty of success or failure, and assuming that this point is slightly above his last previous performance, we have set up the following operational definitions for the present study: achievement oriented aspiration—one point above the last previous performance; failure oriented high aspiration—two or more points above the last previous performance; failure oriented low aspiration—a goal at or below the last previous performance. The performances of the subject are operationally classified as follows: Success—reaching or surpassing one's goal; Partial Failure—a performance one point below the goal set; Failure—a performance two or more points below the goal set.

From the responses, two scores are derived: (a) An achievement oriented score designated as the "A score"; this is the total number of achievement-oriented responses. (b) A failure oriented score designated as the "F score"; this is the algebraic sum of the failure oriented responses weighted in the following manner: a high aspiration response following a Success was given a weight of 1; a high aspiration response following a Partial Failure was given a weight of 2; a high aspiration response following a Failure was given a weight of 3. Similarly, a low aspiration following a Failure was assigned a weight of -1; a low aspiration response following a Partial Failure was assigned a weight of -2, and a low

aspiration response following a Success was assigned a weight of -3.

RESULTS

Table 1 presents results of a *t* test comparing the achievement oriented scores of the Hypertensive group with those of the non-hypertensive Arteriosclerotic group. The difference between the means is not statistically significant, although the direction of the difference shows a higher mean for the Hypertensive group.

Table 2 presents the results of a *t* test comparing failure oriented scores of the Hypertensive group with those of the non-hypertensive Arteriosclerotic group. The difference between the means was significant beyond the .05 level. A mean of 0.38 was found for the Hypertensive group, indicating that this group gave high aspiration responses about as often as low aspiration responses. A mean of -7.61 was found for the non-hypertensive Arteriosclerotic group, indicating that this group predominantly gave low aspiration responses.

Age has not generally been shown to be significantly related to level of aspiration. However, since it is known medically that it is related to arteriosclerosis, the question was raised as to the possibility that age had affected the results. However, analysis revealed identical means for the groups. The average age of the Arteriosclerotic group was 54.17, with a standard deviation of 8.43; the average age for the Hypertensive group was 54.13, with a standard deviation of 8.94. The ages ranged from 32 to 63 in the Arteriosclerotic group and from 34 to 65 in the Hypertensive group. The age factor, then, is not significant in the present study.

TABLE 2

DIFFERENCE BETWEEN THE MEANS OF THE HYPERTENSIVE AND ARTERIOSCLEROTIC GROUPS USING THE F SCORES

Group	N	M	SD	SE _D	t
Hypertensive	24	+0.38	12.45	3.54	2.26**
Arteriosclerotic	23	-7.61	11.79		

** $p < .05$.

Table 3 presents a comparison of the performance scores of the two groups. The mean of the Arteriosclerotic group was 13.91 with a standard deviation of 2.61; the mean of the Hypertensive group was 12.88 with a Standard deviation of 2.04. The difference of 1.03 in the means was not statistically significant.

Table 4 presents a comparison of the degree of improvement of the two groups as measured by the difference between the score on the first trial and the highest score attained on any of the subsequent trials. For the Arteriosclerotic group, the mean improvement was 3.48 with a standard deviation of 1.24. The mean for the Hypertensive group was 3.29 with a standard deviation of 1.28. The slight difference in the mean improvement (.19) was not statistically significant.

DISCUSSION

The intent of this study was to determine whether a group of hypertensive cardiac patients would differ from a group of non-hypertensive arteriosclerotic cardiac patients on a level of aspiration test. The results show that when the level of aspiration is measured in terms of failure oriented responses, the two groups differ significantly with the hypertensives showing a higher level of aspiration response. In Atkinson's (1957) terms there are two ways of responding to fear of failure: choosing a goal in excess of reasonable expectation, or choosing a goal which can be easily attained. Thus, to defend against anxiety, the failure oriented individual either chooses a goal so high that failure to attain the goal is not destructive, or chooses a goal which is low enough so that he can be sure of achieving it. In terms of fear of failure, our data, therefore, suggest that in avoid-

TABLE 3
PERFORMANCE SCORES OF THE HYPERTENSIVE
AND ARTERIOSCLEROTIC GROUPS

Group	N	M	SD	SE _D	t
Hypertensive	24	12.88	2.04	0.65	1.57
Arteriosclerotic	23	13.91	2.61		

TABLE 4

DEGREE OF IMPROVEMENT OF THE HYPERTENSIVE AND ARTERIOSCLEROTIC GROUPS

Group	N	M	SD	SE _D	t
Hypertensive	24	3.29	1.28	0.38	0.50
Arteriosclerotic	23	3.48	1.24		

ing failure, the arteriosclerotic is more likely than the hypertensive to utilize the defense of choosing an easily attainable goal.

Analysis of the frequency of the achievement oriented responses fails to show a statistically significant difference between the groups. However, the significant finding is that when the failure oriented responses of both groups are analyzed, the hypertensive was more likely than the arteriosclerotic to choose a response which was so high as to assure failure. The difference between the arteriosclerotic and the hypertensive lies in the results of their different approaches to failure. The arteriosclerotic avoids the personal feeling of failure by "arranging" for objective success, i.e., by setting his aspiration low. The hypertensive, by contrast, arranges for objective failure. Thus no matter what the reason, the hypertensive less often reaches his goal—that is, he objectively fails.

The results of this study lead us to speculate about a possible relationship between hypertension and continual self-chosen failure. Reiser, Brust, and Ferris (1951) have discussed the role of "life stress" in the development of hypertension, assuming that the patient's reaction to the stress is elevated blood pressure. But just what, in the life stress of the hypertensive, would lead to the blood pressure elevation? Further, what in the life stress distinguishes the hypertensive from the arteriosclerotic, for the psychosomatic literature (Miles et al., 1954) includes studies which find that all heart patients live under more psychological stress than normals.

Our data would support the hypothesis that continued failure, resulting from too high goals, rather than a general life stress, can lead to hypertensive reactions. In at-

tempting to explain this relationship, we hypothesize that continued failure is in some way associated with a physiological withdrawal of blood from the arteriolar bed. The increased blood pressure of the hypertensive follows from the peripheral resistance in the arteriolar bed. Thus the hypertensive reaction is not the direct result of general psychological stress, but rather, is the direct result of the hypertensive avoidance reaction to the specific stress of continued failure. Support for the notion that the hypertensive reaction, i.e., constriction of peripheral vessels, is not due to general psychological stress itself comes from a well-controlled study by Baker and Taylor (1954) which found that skin temperature rises, i.e., arterioles dilate, under general psychological stress.

The present results also force some modification of Dunbar's characterization of the hypertensive. Hypertensives may fear falling short, as she suggests, but they react to this fear, not by choosing goals too low, as she suggests, but by choosing goals so high that they cannot possibly be achieved. Thus they insure failure.

Because this is the first study, to our knowledge, to apply level of aspiration techniques to the study of psychological variables in heart disease, the results obtained suggest further study of cardiac patients with this technique.

SUMMARY

Twenty-four hypertensive cardiac patients and 23 nonhypertensive arteriosclerotic cardiac patients hospitalized at a Veterans Administration Hospital were administered a level of aspiration task based on the Minnesota Rate of Manipulation Test. Achievement oriented scores and failure oriented scores were derived from the aspiration responses. An achievement oriented response was operationally defined as an aspiration one point above the last previous performance; a failure oriented response was defined as either "high" (two or more points above last previous performance), or "low" (at or below the last previous performance). No significant difference between the groups was found in the frequency of achievement oriented scores. However, when failure

oriented responses were broken down into high and low response patterns, the hypertensive group gave significantly ($p < .05$) less low aspiration responses than the non-hypertensive arteriosclerotic group. Level of performance and degree of improvement were not significantly different for the two groups. The hypertensive group "arranged" for repeated failure by consistently setting excessively high goals. It is hypothesized that the withdrawal of blood from the arteriolar bed, resulting in increased blood pressure, is an avoidance reaction to the repeated and continual failure experiences which the failure oriented hypertensive arranges for himself in a neurotic fashion.

REFERENCES

- ADAMS, D. K. Age, race, and responsiveness of levels of aspiration to success and failure. *Psychol. Bull.*, 1939, 36, 573. (Abstract)
- ALEXANDER, F. *Psychosomatic medicine*. New York: Norton, 1950.
- ATKINSON, J. W. Motivational determinants of risk-taking behavior. *Psychol. Rev.*, 1957, 64, 359-372.
- BAKER, L. M., & TAYLOR, W. M. The relationship under stress between changes in skin temperature, electrical skin resistance, and pulse rate. *J. exp. Psychol.*, 1954, 48, 361-366.
- BARNETT, G. J., HANDELSMAN, I., STEWART, L. H., & SUPER, D. E. The occupational level scale as a measure of drive. *Psychol. Monogr.*, 1952, 66(10, Whole No. 342).
- BERKELEY, A. W. Level of aspiration in relation to adrenal cortical activity and the concept of stress. *J. comp. physiol. Psychol.*, 1952, 45, 443-449.
- DUNBAR, F. *Psychosomatic diagnosis*. New York: Harper, 1948.
- FRANK, J. D. Recent studies of the level of aspiration. *Psychol. Bull.*, 1941, 38, 218-225.
- GERARD, D. L., & PHILLIPS, L. Relation of social attainment to psychological and adrenocortical reactions to stress. *AMA Arch. Neurol. Psychiat.*, 1953, 69, 350-354.
- HECHT, I. The difference in goal striving behavior between peptic ulcer and ulcerative colitis patients as evaluated by psychological techniques. *J. clin. Psychol.*, 1952, 8, 262-265.
- IRWIN, F. W., & MINTZER, M. G. Effect of differences in instructions and motivation upon measures of the level of aspiration. *Amer. J. Psychol.*, 1942, 55, 400-406.
- LEWIN, K. *Resolving social conflicts*. New York: Harper, 1948.
- LEWIN, K., DEMBO, TAMARA, FESTINGER, L., & SEARS, PAULINE. Level of aspiration. In J. McV. Hunt (Ed.), *Personality and the behavior disorders*. Vol. 1. New York: Ronald, 1944. Pp. 333-378.

- LITTLE, SUE W., & COHEN, L. D. Goal setting behavior of asthmatic children and of their mothers for them. *J. Pers.*, 1951, 19, 376-389.
- MILES, H. H. W., WALDFOGEL, S., BARRABEE, EDNA L., & COBB, S. Psychosomatic study of 46 young men with coronary artery disease. *Psychosom. Med.*, 1954, 16, 455-477.
- RAIFMAN, I. Level of aspiration in a group of peptic ulcer patients. *J. consult. Psychol.*, 1957, 21, 229-231.
- REISER, M. F., BRUST, A. A., & FERRIS, E. B., JR. Life situations, emotions and the course of patients with arterial hypertension. *Psychosom. Med.*, 1951, 13, 133.
- REISSMAN, L. Levels of aspiration and social class. *Amer. sociol. Rev.*, 1953, 18, 233-242.
- SAJI, M. The degree of reality in level of aspiration. *Jap. J. Psychol.*, 1951, 21(3-4), 56-69.
- SCODELL, A. Passivity in a class of peptic ulcer patients. *Psychol. Monogr.*, 1953, 67(10, Whole No. 360).
- SEARS, P. S. Levels of aspiration in academically successful and unsuccessful children. *J. abnorm. soc. Psychol.*, 1940, 35, 498-536.
- STUBBINS, J. The relationship between level of vocational aspiration and certain personal data: A study of some traits and influences bearing on the prestige level of vocational choice. *Genet. psychol. Monogr.*, 1950, 41, 327-408.
- WALDER, L. O. The effects of instructions, order, and sex-ethnic group in level of aspiration situations. Unpublished masters dissertation, University of Hawaii, 1951.
- WALTER, L. M. The relation of age and sex to levels of aspiration. Unpublished masters dissertation, Illinois State Normal University, 1948.
- WALTER, L. M., & MARZOLF, S. S. The relation of sex, age, and school achievement to levels of aspiration. *J. educ. Psychol.*, 1951, 42, 285-292.

(Received July 20, 1960)

OBJECTIVE ESTIMATES OF CLINICAL JUDGMENTS¹

ROLFE LAFORGE

University of Illinois

One of the more promising avenues towards a theory of personality develops from the discovery of the kind of structure most appropriate for a model of clinical inference. The present paper reports an early and imperfect example of this approach. There are at least four major imperfections.

First, while one would prefer to abstract representative tasks of clinical inference, in this study only one kind of information and only one clinical construct were available. MMPI profiles of scores on the 12 *K*-corrected clinical scales were given to four psychologists experienced in the use of the MMPI. Each psychologist rated the "degree to which *repression* was relied upon by the patient as a defense against anxiety." The average of these four ratings for any patient is the value of the "clinically inferred" variable (*R*).²

Second, one would like to ask the psychologists who make the clinical inferences to describe also the cues, working hypotheses, and methods of verification which they use. Here, each rater was asked only to list as many "cues" of repression as could be found in MMPI profiles. The four lists were pooled, and duplications removed; 27 cues remained, with many dependencies among them.

Third, one would like to examine the relations between data and construct, and between construct and prediction, using the most general statistical models available. In this study only two models were applied. One model employed a linear prediction of *R* from its multiple regression on *H_y* and *Sc*. The other method sought cues which could be

associated with different ordinal subregions within the total range of *R*. That is, the variable *R* was used to order the profiles within a sample; then the Mann-Whitney and the Wald-Wolfowitz (runs) tests were applied separately to each of the suggested cues to test for relationship to *R*. A 5% level of significance was used in all tests. In the first sample of 35 outpatients, eight cues were related to *R* by the Mann-Whitney. Four of these same cues were related to *R* according to the runs test, plus two additional cues which had a nonmonotonic relationship with *R*. (The four found by the Mann-Whitney alone are presumably the result of its greater power against monotonic alternatives to the null hypothesis.)

Fourth, and most serious of the imperfections, cross-validation of these results on the second sample of 83 outpatients became impossible as a result of loss of data during a 6-year lapse. The investigator redefined from memory 26 cues whose relation to *R* was then tested in the second sample by the runs test alone. Of 14 "pattern" cues referring to the relative height of two or more scales, 11 were found to relate to *R*; of 11 "absolute elevation" cues, 4 related to *R*. One "mixed" cue, referring to both elevation and pattern, also predicted *R*.

These 16 cues could be used as a checklist to enable the MMPI novice or researcher to make a quick estimate of *R* from an MMPI profile. To simplify the estimation process, a Guttman scale was constructed by eliminating cues approximately equal in relative frequency to others. The remaining seven cues scaled in the second sample with a reproducibility above .90; removal of one cue raised this to .96. A check on a third sample of 201 outpatients gave a reproducibility of .93. Be-

¹ This work was supported by United States Public Health Service Postdoctoral Fellowship M3634 (1952).

² For details, see: LaForge, Leary, Naboisek, Coffey, Freedman (1954, pp. 132, 135-136).

cause the cues are to some extent experimentally dependent, this figure indicates only the sufficiency of the estimate in use. The six-item Guttman-type scale³ correlated as well with R (.799) as did the simple sum over all predictive items (.788). In comparison, the multiple correlation of H_y and Sc with R , based on the same sample of 83 cases, was .843. (The beta weights in this sample were, for K -corrected T scores: H_y , .0718; Sc , -.0460. Thus a quick estimate of R could be obtained by subtracting Sc from $1\frac{1}{2}$ times H_y .)

On the basis of the present evidence, there is little to choose between the two estimates. Of course the lack of cross-validation makes

³ The Guttman-scaled cues are $L \geq 60$, $L \geq F$, $H_s + H_y \geq D + Sc$, $H_y \geq Sc$, $K + 10 \geq F$, and $H_y + 10 \geq D$, in order of increasing frequency of occurrence. The most predictive single cue was $H_s + H_y \geq D + Sc$.

comparison particularly difficult. Consideration of those cues which predicted R as against those which did not significantly discriminate suggests that the emphasis on configuration or pattern in the teaching of clinical inference may be justified. On the other hand, the equal success of a linear combination of two scales points up the psychometrician's faith in the efficacy of negatively correlated ($r_{H_y Sc} = -.379$) predictors in an apparently configural situation. And either objective measure showed correlations with R approximately as high as the interrater correlations.

REFERENCE

- LAForge, R., LEARY, T. F., NABOISEK, H., COFFEY, H. S., & FREEDMAN, M. B. The interpersonal dimension of personality: II. An objective study of repression. *J. Pers.*, 1954, 23, 129-153.

(Received June 20, 1960)

A NOTE ON "IMPULSE REPRESSION AND EMOTIONAL ADJUSTMENT"

H. J. EYSENCK

University of London

In a recent paper, Grater (1960) has tested the Freudian theory of impulse repression as a correlate of neuroticism, emerging with conclusions which apparently contradict that theory; the more neurotic subjects were, if anything, less "repressive" than were the non-neurotic ones. This result would appear to be in line more with Mowrer's (1953) view of neurosis, according to which

the problem-solving activity which is usually referred to clinically as self-protectiveness or defensiveness . . . functions in the interest of the primary drives or id, rather than, as Freud posited, in the services of the socially derived forces of the superego (p. 145).

I have discussed the point at issue between orthodox Freudian writers and Mowrer elsewhere (Eysenck, 1957, p. 82f.), and have suggested there that the distinction which must be made in order to accommodate the known facts is one between *extraverted* neurotic behavior patterns (hysteria, psychopathy, hypochondriasis, etc.) and *introverted* neurotic behavior patterns (anxiety, reactive depression, obsessional-compulsive, dysthymic reactions). This personality dimension of extraversion-introversion is conceived of as being orthogonal to neuroticism, and I have further suggested that "impulse repression" and socialization generally are in part caused by constitutional factors closely linked with introversion. Dysthymic neurotics, according to this view, are "oversocialized," hysteric and psychopathic ones "undersocialized." This theory has been discussed in some detail in relation to the experimental evidence (Eysenck, 1957, 1960a, 1960c) and it may be concluded that it serves to reconcile a large amount of factual material.

When we turn to Grater's study we find that he has defined neuroticism in terms of

three MMPI scales two of which are measures of extraverted neuroticism (*Hy*, *Hs*), while the most clear-cut introverted scale (*Pt*) was not used at all. According to the analysis given above, therefore, we would expect neurotics (as defined by Grater's MMPI scores) to be extraverted and less given to impulse repression than nonneurotics. His results, as far as they go, bear out this prediction, although in only one or two instances do his scores reach statistical significance.

The purpose of this note is not so much to reinterpret Grater's data as to draw attention to the absolute necessity, in work of this kind, to take into account the two-dimensional nature of the test-space in which the experimenter is working (Eysenck, 1960b). Much experimental work in this field is wasted because results are quite uninterpretable, it being impossible from the data given to sort out the dimensions involved; work with the Manifest Anxiety scale is a good example of this, the resulting score having loadings both on neuroticism and on introversion (Bendig, 1960; Eysenck, 1957). Much of the theoretical disputation regarding the nature of neuroticism is sidetracked by emphasizing either the extraverted or the introverted side (Miller & Dollard, 1950; Mowrer, 1953). The evidence for *at least* two factors in this field is now practically conclusive (Eysenck, 1960b) and it would seem desirable to recognize this fact in the design and interpretation of psychological experiments.

REFERENCES

- BENDIG, A. W. Factor analysis of "anxiety" and "neuroticism" inventories. *J. consult. Psychol.*, 1960, 24, 161-168.
EYSENCK, H. J. *The dynamics of anxiety and hysteria*. London: Routledge & Kegan Paul, 1957.

- EYSENCK, H. J. *Experiments in personality*. London: Routledge & Kegan Paul, 1960. 2 vols. (a)
- EYSENCK, H. J. *The structure of human personality*. (2nd ed.) London: Methuen, 1960. (b)
- EYSENCK, H. J. Symposium: The development of moral values in children: VII. The contribution of learning theory. *Brit. J. educ. Psychol.*, 1960, 30, 11-21. (c)
- GRATER, H. A. Impulse repression and emotional adjustment. *J. consult. Psychol.*, 1960, 24, 144-149.
- MILLER, N. E., & DOLLARD, J. *Personality and psychotherapy*. New York: McGraw-Hill, 1950.
- MOWRER, O. H. *Psychotherapy: Theory and research*. New York: Ronald, 1953.

(Received July 1, 1960)

SEX DIFFERENCES IN MENTAL HEALTH ANALYSIS SCORES OF ELEMENTARY PUPILS

JOSEPH C. BLEDSOE

University of Georgia

The manner in which the individual perceives himself has been regarded as one indicator of the degree of mental health he demonstrates or possesses. Among the currently available instruments for measuring self-perception of mental health status is the Mental Health Analysis prepared by the California Test Bureau. The test manual makes no mention of sex differences in norms for the Mental Health Analysis. Previous studies (Ausubel, Balhazar, Rosenthal, Blackman, Schpoont, & Welkowitz, 1955; Davidson, Sarason, Lighthall, Waite, & Sarnoff, 1958; Sarason, Davidson, Lighthall, & Waite, 1958) have indicated that girls may perceive themselves as significantly more accepted and intrinsically valued than boys. The present study provides more evidence of the possibility of sex differences in mental health status, as reflected by responses to the Mental Health Analysis.

The subjects were 96 girls and 101 boys enrolled in the fourth through the seventh grades in an elementary school of a southeastern city of approximately 35,000 population. The Elementary Form of the Mental Health Analysis was administered as a part of a 3-year in-service education program for teachers designed to promote better understanding of mental health principles and practices.

The Mental Health Analysis consists of 200 items to be answered by circling Yes or No. Five sorts of personality "liabilities" and five sorts of "assets" are investigated. Liabilities include behavioral immaturity, emotional instability, feelings of inadequacy, physical defects, and nervous manifestations. The assets include close personal relationships, interpersonal skills, social participation, satisfying work and recreation, and adequate outlook

and goals. The reliability of the Elementary scale is reported as: assets, .90; liabilities, .89; total, .90. Reliability coefficients for component scores of the Elementary scale vary from .80 for "outlook and goals" to .85 for "physical defects." Content validity of the instrument is based upon the adequacy of item selection, the meaningfulness of the analysis of the mental health categories, and the cleverness in disguise of items. Studies of concurrent validity are reported in the test manual (California Test Bureau, 1959).

Means and standard deviations for the several components, categories, and total scores on the Mental Health Analysis for boys and girls were computed.¹ Differences in means were then tested for significance by the *t* test of significance. The .05 level of confidence was accepted as the criterion for rejection of the null hypotheses involved. Twelve of the 13 differences favored the girls, and 7 of these differences met the .05 level of significance criterion. Significant differences were found in the total score, the total assets, the total liabilities, and in close personal relationships, outlook and goals (assets subscales), and behavioral immaturity and feelings of inadequacy (liabilities subscales). A nonsignificant difference favoring the boys was found in the physical defects subscale of the liabilities category. Thus within the limitations of the study, it appears that elementary school girls tend to rate themselves significantly higher on

¹ A summary table of these statistics has been deposited with the American Documentation Institute. Order Document No. 6762 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance \$1.25 for microfilm or \$1.25 for photocopies. Make checks payable to: Chief, Photoduplication Service, Library of Congress.

the Mental Health Analysis than do elementary age boys. Implications for differential norms may be suggested. More important may be implications for teacher understanding of and curriculum adjustment to the needs of boys.

REFERENCES

- AUSUBEL, D. P., BALTHAZAR, E. E., ROSENTHAL, I., BLACKMAN, L. S., SCHPOONT, S. N., & WELKOWITZ, J. Perceived parent attitudes as determinants of children's ego structure. *Child Develpm.*, 1955, 25, 173-183.
- CALIFORNIA TEST BUREAU. *Manual for Mental Health Analysis*. (Rev. ed.) Los Angeles: CTB, 1959.
- DAVIDSON, K. S., SARASON, S. B., LIGHTHALL, F. F., WAITE, R. R., & SARNOFF, I. Differences between mothers' and fathers' ratings of low anxious and high anxious children. *Child Develpm.*, 1958, 29, 155-160.
- SARASON, S. B., DAVIDSON, K., LIGHTHALL, F., & WAITE, R. Rorschach behavior and performance of high and low anxious children. *Child Develpm.*, 1958, 20, 277-285.

(Early publication received December 12, 1960)

BRIEF REPORTS

A COMPARISON OF ACCEPTORS AND RESISTORS OF DRUG TREATMENT AS AN ADJUNCT TO PSYCHOTHERAPY¹

ALLEN RASKIN

Veterans Administration, Washington, D. C.

In a recent adjunct chemotherapy study with psychiatric outpatients, two of the major causes of patient attrition were refusal to take the study medication and excessive deviation from prescribed dosage levels. Can these patients be identified early in treatment and why are they reluctant to take the assigned medication? This study was undertaken as a preliminary effort to answer these questions by identifying variables which differentiate patients who remained in psychotherapy but resisted taking drugs from patients who accepted drug treatment as an adjunct to psychotherapy.

Data for the present study were collected in 22 Veterans Administration Mental Hygiene Clinics as part of a larger chemotherapy study. There were four psychotherapy-plus-drug groups and one psychotherapy-only group in the larger study. The four study drugs were chlorpromazine, meprobamate, phenobarbital, and placebo. The study was conducted double-blind. Patients were told that the drugs had helped a lot of people with similar troubles. Study patients were all males, under age 50, who were acceptable for individual psychotherapy. There were 142 patients (the Acceptors) who remained in psychotherapy at least 8 weeks and took their medication as prescribed. An additional 37 patients (the Resisters) also remained in psychotherapy for at least 8 weeks but either refused to take the as-

signed medication or took significantly less than the amount prescribed. As Acceptors and Resisters did not differ significantly by treatment group, all Acceptors were pooled into one group and all Resisters into another.

On the basis of data obtained from the patients and from their therapists, the Acceptors and Resisters were compared on 40 personality, socioeconomic, and attitudinal variables. Resisters did not report a greater number of adverse side effects at the end of the first week on medication. Eleven significant differences were found between these two groups as compared to two expected by chance for the number of tests made. Compared to the Acceptors, the Resisters were better educated, had a greater knowledge of psychiatry, had less favorable attitudes toward physicians, rated themselves more hostile on an adjective checklist, and admitted to greater direct expression of hostility. At the end of the initial interview, therapists rated Resisters as expressing a more negative attitude toward taking the assigned drugs, less likely to have their psychotherapy facilitated by the addition of a tranquilizer, more inwardly hostile, reporting more incidents of overt aggression, and less likable than Acceptors. After 8 weeks of treatment, therapists also rated the Resisters as more resistive to psychotherapy.

These findings indicate that by the end of the initial psychotherapy session, therapists were able to identify many Resisters by their negative attitude toward taking the assigned medication. The Resisters' reluctance to take the study drugs was not due to side effects of the drugs. Instead, the drugs apparently provided a convenient focal point for the hostile and aggressive impulses noted by the Resisters' therapists and admitted by the patients themselves.

(Received September 23, 1960)

¹ An extended report of this study may be obtained without charge from Allen Raskin (Neuropsychiatric Research Laboratory, Veterans Benefits Office; Munitions Building; Washington 25, D. C.) or for a fee from the American Documentation Institute. Order Document No. 6647 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance \$1.25 for microfilm or \$1.25 for photocopies. Make checks payable to: Chief, Photoduplication Service, Library of Congress.

AN INTERPRETATION OF m ¹

JOHN M. REISMAN

Rochester Child Guidance Clinic, New York

Piotrowski (1957) hypothesized that the types of movement in m responses are always different from those of M and FM . His hypothesis was based on his interpretations of m as tendencies which are never acted out while tendencies indicated by M and FM might be manifested in overt behavior. These "tendencies" were said to be primarily indicated by three types of action: flexor, extensor, and blocked movements. Respectively, they were characterized by giving in to gravity, overcoming gravity, and tension; they were interpreted as tendencies toward compliance, self-assertion, and indecisiveness.

Using a sample of disturbed children whose mean age was 10, Reisman (1960) found results directly opposite to those predicted by the hypothesis. This study repeated the test of the hypothesis with adolescents to determine if, with increasing age, there is a tendency for a person's m to differ in type from his M and FM .

Only responses that were identical with or highly similar to Piotrowski's examples of extensor, flexor, and blocked movements were considered. With this restriction, from a pool of 80 Rorschachs obtained from adolescents referred to a child guidance clinic, only 22 were found which contained at least 1 m and 1 M . In this sample, there were 18 boys and 4 girls. Ages ranged from 13-15; IQ scores ranged from 80-120 ($\bar{X} = 102$). None of the subjects was considered psychotic. All but one of the subjects, who was referred for underachieving in school, had been referred for acts of delinquency.

In a test for reliability, there was 90% and 80% agreement between the experimenter and

two judges in categorizing 96 responses as to type. Disagreements did not appreciably affect the results. Movement responses were extracted from their records, coded, and written verbatim in a random order. Three months elapsed between the compilation and the categorizing of responses.

Subjects produced a scorable total of 29 extensor, 19 flexor, and 2 blocked M responses; 30 extensor, 6 flexor, and 0 blocked FM responses; and 25 extensor, 2 flexor, and 1 blocked m responses. These results were similar to those of the previous study and to findings with a sample of "normal" adolescents. Of the 22 records, 18 had m and M or FM responses of the same type; 14 records had m and M responses of the same type. These results were, once again, in direct opposition to Piotrowski's hypothesis. Furthermore, if it had been predicted that an adolescent's M or FM is of the same type as his m , this hypothesis would have been supported ($\chi^2 = 7.68$; $p < .01$).

Piotrowski has stated that tendencies indicated by m are never acted out. In only one case, that of a boy whose m was extensor and who was referred for underachieving, was this expectation supported by the record of overt behavior. It would seem reasonable to conclude that, depending upon the adolescent's controls, the tendencies indicated by m may or may not be acted out. The results further suggest that m be interpreted as an awareness of impulses or feelings which the individual experiences difficulty in controlling and expressing. Piotrowski's types of movement seem to offer assistance in understanding what kinds of impulses or feelings cause the person concern.

REFERENCES

- PIOTROWSKI, Z. N. *Perceptanalysis*. New York: Macmillan, 1957.
REISMAN, J. M. Types of movement in children's Rorschachs. *J. proj. Tech.*, 1960, 24, 46-48.

(Received October 7, 1960)

¹ An extended report of this study may be obtained without charge from John Reisman (Rochester Child Guidance Clinic; 31 Gibbs Street; Rochester 4, New York) or for a fee from the American Documentation Institute. Order Document No. 6648 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance \$1.25 for microfilm or \$1.25 for photocopies. Make checks payable to: Chief, Photoduplication Service, Library of Congress.

THE RELATION OF THE FAMOUS SAYINGS TEST TO SELF- AND IDEAL-SELF-ADJUSTMENT¹

BERNARD I. MURSTEIN

Interfaith Counseling Center, Portland, Oregon

Bass (1958, p. 479) cites Thurstone as stating "the best form of projective test is one which is quite unstructured for the subject but fairly well structured for the examiner." Using Thurstone's words as a model, Bass attempted to construct a new projective technique in which administration and scoring are completely objective. The subject is presented with a booklet of 100 proverbs which he answers by checking the Yes, ?, or No box for each proverb. From his answers four slightly correlated scales have been derived. These are Social Acquiescence, Fear of Failure, Conventional Mores, and Hostility. Social Acquiescence has received wide publicity because of the role it is reputed to play in most personality questionnaires. Little, however, has been done to study the meaning of Social Acquiescence apart from its role as an interferer in the study of other variables. Accordingly, to study the personality correlates of the Famous Sayings Test (FST), scores on this test were compared to a personality questionnaire adaptation of the Butler-Haigh SIO Q sort instrument.

The hypothesis for this study was that there is no significant correlation between the scales of the FST and the following scales of the SIO Q sort: (a) Self-Adjustment, (b) Ideal-Self-Adjustment, and (c) Self-Ideal-Self-Adjustment Discrepancy.

Forty-five subjects (24M, 21F) from the introductory psychology class at the University of Portland were told that the experimenter wished to investigate the merits of a new type of questionnaire about how people describe themselves as

well as to construct a test of attitudes towards famous sayings.

Of 15 correlations bearing on the hypothesis, only the correlation between Social Acquiescence and Self-Ideal-Self-Adjustment Discrepancy proved to be significant at the .01 level ($r = .45$). The results lend support to the interpretation of Social Acquiescence as more than an interfering set. Social Acquiescence may be a superficial phenotype reflecting considerable underlying anxiety which manifests itself by an inordinate need to win acceptance by others through excessive conformity.

As to the other scales—Conventional Mores, Hostility, and Fear of Failure—none showed any relation to Self- or Ideal-Self-Adjustment. Considering earlier failures to find meaningful correlates for these scales, it must be considered doubtful that they will eventually find a niche for themselves in trait measurement. The reason for this pessimism is that, in addition to rather low reliabilities, these scales make the further, probably unwarranted, assumption that saying "yes" to a hostile proverb indicates that one is hostile. It must be remembered that proverbs are not necessarily simple, bland statements that have no consensual validation about their general truth. Accordingly, it is entirely possible that many of the proverbs may be perceived as being true by the majority of persons. Hence, there can be little differential value as to these scales unless scaled values can be established as to the normative agreement on the verification of a proverb. To a lesser degree the same criticism is applicable to Social Acquiescence. The fact that it contains 56 items compared to 30 for each of the others and is not limited to a specific content as is true of the others, probably accounted for its greater validity.

REFERENCE

- BASS, B. M. Famous Sayings Test: General manual. *Psychol. Rep.*, 1958, 4, 479-497.

(Received October 17, 1960)

¹ An extended report of this study may be obtained without charge from Bernard Murstein (Interfaith Counseling Center; 729 Southwest Alder; Portland, Oregon) or for a fee from the American Documentation Institute. Order Document No. 6759 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance \$1.25 for microfilm or \$1.25 for photocopies. Make checks payable to: Chief, Photoduplication Service, Library of Congress.

THE INTELLIGENCE OF BOYS WITH MUSCULAR DYSTROPHY¹

DON KEITH WORDEN

Western Reserve University

Since pseudohypertrophic muscular dystrophy (PMD) was first described there has been controversy concerning the intelligence of the children who have it. Several studies done recently claim that these children do not suffer from mental deterioration, and that they are of average intelligence. Academic retardation has been mentioned as a frequent problem, in some of these studies.

This study was designed to clarify some of these points in a more rigorous manner than has been done to date. A group of boys with PMD was given intelligence and achievement tests.

Several further questions were raised in this investigation: (a) Would the measured depression in intelligence be partly a function of the child's socioeconomic background? (b) Does any child with a chronic disease show depression in intelligence? (c) Would measured depression in intelligence be due partly to the factor of the child's having a physically handicapping (as distinct from merely chronic) illness? Appropriate control groups were selected to investigate these questions for: (a) siblings of the PMD group, (b) diabetics, and (c) amyotonia congenita.

Form M of the 1937 Stanford-Binet scale was used as the measure of intelligence. The measures of academic achievement were the Gilmore Oral Reading Test and the Metropolitan Achievement Test in Arithmetic Fundamentals. The educational quotient ($EA/CA \times 100$) and the accom-

plishment quotient ($EA/MA \times 100$) were figured.

The muscular dystrophy group consisted of 38 school age boys. It was possible to test 27 of these children's siblings on the Binet.

The amyotonia congenita group consisted of 16 children. The diabetic and the diabetic-sibling group consisted of 36 children each.

The PMD group had a mean IQ of 83. The range was from 46 to 134, but skewed radically to the left.

Twenty-five of these 38 boys were given the achievement test in arithmetic. The mean educational quotient (EQ) of this group was 84. The accomplishment quotient (AQ) was 96. The reading test was administered to 24 of these 38 boys. The mean educational quotient for reading was 87 and the mean accomplishment quotient was 101. The IQ did not differ significantly from the EQ in either reading or arithmetic. These three measures, however, differed significantly from the AQ in reading and the AQ in arithmetic. The two AQs did not differ significantly from one another.

The siblings of the PMD group had a mean IQ of 110. The children with diabetes mellitus had a mean IQ of 107. The diabetic siblings had a mean IQ of 109. The patients with amyotonia congenita had a mean IQ of 118. The analysis indicated that the PMD group differed significantly from all control groups.

In summary, boys with pseudohypertrophic muscular dystrophy functioned on an intelligence test significantly below average and below several control groups. Their IQs and EQs are both below average and do not differ from one another. However their AQs indicate that they are doing all one could expect for children of their mental ages.

Thus, within the confines of this study, the intellectual deficit seems specifically associated with the factor of having PMD.

(Received October 31, 1960)

¹ An extended report of this study may be obtained without charge from Don K. Worden (Southeastern Ohio Guidance Center; Athens, Ohio) or for a fee from the American Documentation Institute. Order Document No. 6760 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance \$1.75 for microfilm or \$2.50 for photocopies. Make checks payable to: Chief, Photoduplication Service, Library of Congress.

SYMBOLIC INTERPRETATION OF RORSCHACH CONTENT¹

JOSEPH F. RYCHLAK

St. Louis University

AND

DONALD E. GUINOUARD

Montana State College

This research continues the examination of 12 Rorschach contents studied earlier (Rychlak, 1959): Bat, Bear, Boots, Clouds, Fire, Fur, Hair, Island, Mask, Mountains, Rocks, and Smoke. It was decided that a culling of Rorschach protocols for these contents, and then a comparison made along personality dimensions for subjects perceiving certain of these contents might throw preliminary light on the little studied area of content symbolism.

Subjects were 80 girls and 86 boys in the 11-14 year age range. Testing was done in groups, with the Harrower (1959) Group Rorschach slides and the High School Personality Questionnaire (Cattell, Beloff, & Coan, 1958) used as inkblot and personality measures, respectively. All testing was completed within one week. Scoring percentage of agreement between judges for identification of the contents was 95; subjects' one week test-retest reliabilities—based on binomial expansion—reached the .01 level for half of the contents.

Frequently reported contents for all subjects ($N = 166$) included Bat, Bear, and Fire; whereas Fur, Hair, and Island were infrequently noted. Girls reported Clouds, Hair, and Smoke content significantly more often than boys ($p < .05$ or less). Boys were more likely than girls to report Mask content ($p < .01$).

The (chi square) .01 level personality findings for the total sample were: Subjects who see Boots tend to be talkative and cheerful. Children who report Bat content are less tense and excitable than their peers. The reporting of Fur is related to a form of emotional instability, suggest-

ing overactivity and frustration. The .05 level findings for total sample were: Fire content is reported more frequently by subjects who are tense and excitable. Children who report Rocks content tend to be dominant, independent, and outgoing. Island content is suggestive of a tough, self-sufficient personality.

Considered individually by sex, the following .05 level findings were noted: Boys who see Clouds in the inkblots are phlegmatic, deliberate, and self-effacing; girls who see Clouds are tough, realistic, and group-conforming. Girls reporting Hair are self-concerned and individualistic. The contents Bear, Smoke, Mask, and Mountains did not discriminate between subjects' personalities.

To propose to study potential symbols nomothetically implies the existence of a kind of "universal" symbolism, something repugnant to the common sense of many clinicians. Yet, in broad terms, such a symbolism may be possible through the intervention of cultural factors. Reference to a collective unconscious by some theorists would be interpreted from this viewpoint to mean that there are variables in any culture which—though learned by all or by many—are un verbalized. They are the shadow side of cultural manifestation. The clinician, through studies of this sort and the assumption of the proper set as a result of his study, may identify these variables in his client's projective responses, dreams, etc. In this connection, interesting parallels were noted between the present findings and the forced associations of the earlier study (Rychlak, 1959).

REFERENCES

- CATTELL, R. B., BELOFF, H., & COAN, R. W. *High School Personality Questionnaire*. Champaign, Ill.: Institute for Personality & Ability Testing, 1958.
- HARROWER, MOLLY R. *Group Rorschach slides*. New York: Psychological Corporation, 1959.
- RYCHLAK, J. F. Forced associations, symbolism, and Rorschach constructs. *J. consult. Psychol.*, 1959, 23, 455-460.

(Received November 23, 1960)

CLIENT DEPENDENCY AND THERAPIST EXPECTANCY AS RELATIONSHIP MAINTAINING VARIABLES IN PSYCHOTHERAPY

KENNETH HELLER

University of North Dakota

AND

ARNOLD P. GOLDSTEIN

University of Pittsburgh School of Medicine

The relationship and interactional aspects of psychotherapy, while long considered important, have received ever-increasing theoretical and experimental emphasis in recent years. Bordin (1959), for example, has stated:

The key to the influence of psychotherapy on the patient is in his relationship with the therapist. Wherever psychotherapy is accepted as a significant enterprise, this statement is so widely subscribed to as to become trite. Virtually all efforts to theorize about psychotherapy are intended to describe and explain what attributes of the interactions between the therapist and the patient will account for whatever behavior change results (p. 235).

Previous investigations of the psychotherapeutic relationship have frequently been attempts to identify and interrelate client, therapist, and/or transactional variables which theoretically appear to be important dimensions of the therapist-client interaction. The present study, continuing in this direction, has focused upon client-therapist attraction as its major concern. This potentially significant dimension of the therapeutic relationship has been defined by Libo (1957), in a general sense, as "the resultant of all forces acting on the patient to maintain his relationship with the therapist." To implement this definition, he developed the Picture Impressions Test, a projective technique designed to call forth client verbalizations concerning feelings toward therapists and the therapy process. With client-therapist attraction defined in this manner, Libo (1957) demonstrated a significant relation between the magnitude of client attraction toward the therapist and certain clearly observable and therapy relevant, overt, client behaviors.

There is evidence to suggest that the specific nature of these relationship maintaining forces acting upon the client may include such

participant characteristics as client dependency and the therapist's expectations regarding a favorable therapeutic outcome. In a discussion of dependency in psychotherapy, Dollard and Miller (1950) note that therapy is often facilitated by initial client dependency. According to their formulation, the client brings to the therapeutic situation a desire to please the therapist, this desire being considered one of the main forces helping the client overcome the initial anxieties associated with therapy. As therapy progresses the client is expected to grow in independence, since he need no longer rely on pleasing the therapist as his only motivation for continuing in therapy.

The viewpoint that dependent clients become more independent after the successful completion of psychotherapy is also examined by another line of investigation. Studying the present-self and ideal-self descriptions of psychiatric patients, Fordyce (1953) found that those patients who described themselves as dependent stated that they would ideally like to see themselves as being more independent. Since Rogers and Dymond (1954) and others report that successful psychotherapy produces an increased congruence between present-self and ideal-self descriptions, it seems reasonable to expect that, after a course of successful psychotherapy, clients with pretherapy dependent self-descriptions should see themselves as growing in independence.

The therapist's expectation of patient improvement, a second potentially important relationship maintaining variable, has been demonstrated by Goldstein (1960b) to affect significantly the amount of improvement the patient reports as having taken place and also the duration of psychotherapy. Kelley (1949), Rosenthal (1959), and Ulenhuth,

Canter, Neustadt, and Payson (1959) have also demonstrated the potency of participant expectancies in two-person interactions.

Hypotheses

1. Client pretherapy attraction to the psychotherapist varies: (a) positively with client pretherapy dependency, and (b) positively with client over-therapy movement toward independence.
2. Client pretherapy attraction to the psychotherapist varies positively with the latter's expectation of client improvement.

METHOD

Two treatment conditions were utilized in the present investigation: therapy (experimental group), and no-therapy (control group). Thirty clients and 10 therapists participated. Most of the clients were undergraduates in attendance at the Pennsylvania State University who had sought psychotherapy at the University Psychological Clinic. Clients were randomly assigned to the two treatment conditions and the 15 clients in the experimental group were then randomly assigned to therapists. Each therapist met with his client(s) two times per week for individual, 50-minute sessions. The 15 control clients were placed on a waiting list and did not participate in formal psychotherapy during the 15 session duration of the investigation.

The 10 therapists employed in this study had all completed their formal predoctoral training in psychotherapy, including an approved internship. Eight therapists had yet to fulfill all the requirements for the PhD degree, while two had already received their PhD degrees. The therapists, all males, ranged in age from 24 to 37, with a median age of 32. When asked to describe their own orientation to therapy, none of the therapists expressed a strong preference for any particular "school" of psychotherapy. All stated that their approach would vary according to the client and the situation. The therapists were employed by either the Psychological Clinic or the Division of Counseling at Pennsylvania State University and usually saw clients as part of their regular clinical case load.

Measurement of Variables

Client-therapist attraction. As suggested above, client-therapist attraction was operationally defined as the client's score on the Picture Impressions Test. This projective technique consists of four cards depicting therapy-like situations to which the client is requested to respond in a manner analogous to TAT administration. Content analysis scoring (Libo, 1956) (e.g., Locomotion, Barriers to Locomotion, Satisfaction, etc.) was carried out independently by the authors with complete agreement occurring on 83% of the client stories. For each client, an at-

traction score was then determined by summing his scores for each of his four stories.

Dependent behavior. Dependent behavior was conceived of as the extent to which an individual prefers to have others prevent his frustration or punishment and provide need satisfaction (Fitzgerald, 1958). In order to narrow the definition of dependency even further, it was decided to concentrate on two aspects of dependent behavior described by Murray (1943) as Succorance and Deference. Measurement of dependent behavior occurred at two levels:

Self-descriptive dependency. To measure the extent to which clients attributed dependent behavior to themselves, the Succorance, Deference, and Autonomy scales of the Edwards Personal Preference Schedule (EPPS) (Edwards, 1954) were administered. Following the suggestion of several researchers (Bernardin & Jessor, 1957; Gisvold, 1958; Zuckerman & Grosz, 1958) a total self-descriptive dependency score was computed by summing the scores from the Succorance and Deference scales and subtracting from the sum the score from the Autonomy scale.

Overt dependency. A Situational Test of Dependency developed earlier by one of the authors (Heller, 1959) from a modification by Borgatta (1951) of the Rosenzweig P-F study (1947) was used in the current investigation. Borgatta developed a role-playing form of the P-F study in which the original paper and pencil situations were acted out by both examiner and examinee. Borgatta's evidence suggests that subjects react to the role-playing form of this test in a manner quite similar to the way in which they react to real, overt, threatening situations. The role-playing situations were further modified so that all the situations involved a degree of threat to the respondent. An additional modification was development of a forced-choice rather than an open-ended method of responding.

Therapist expectation. Therapist expectation of client personality change was generally defined as the feelings held by the therapist relating to the anticipated nature and intensity of his client's personality problems upon completion of the latter's psychotherapy. Operationally, this variable was defined (Goldstein, 1960b) as the difference between the therapist's ordering of personality problem Q sorts when he is instructed to sort them under two different orientations: (a) according to the status in which he, the therapist, expects his client's problems to be upon completion of psychotherapy; and (b) according to the manner in which he views his client's problems at the time of sorting, i.e., his present perception of his client.

The Picture Impressions, EPPS, and the Situational Test of Dependency were individually administered to all clients immediately prior to their first therapy session and immediately following their fifteenth session. If a therapy client dropped out of therapy before his fifteenth session, he was "post-tested" at the time of dropout. When this occurred, a control client who had been in the wait group for the same period of time as the experimental client

had been in therapy, was also tested. The therapists completed their sortings after every 5 sessions for 15 sessions.

RESULTS AND DISCUSSION

The correlations, for both experimental and control clients, between pretherapy attraction and the dependency scores obtained pre- and posttherapy, as well as the resultant dependency difference score, are presented in Table 1.

The pretherapy correlations in Table 1 indicate a significant relationship between client's attraction and both self-descriptive and overt dependency. Those individuals who wrote stories to the Picture Impression cards indicating that they anticipated positive gratification from therapy, described themselves before therapy as more dependent according to the EPPS, and also acted more dependently on the Situational Test of Dependency. This finding lends support to the contention of Dollard and Miller that initial client dependence can act in ways that maintain the early stages of the psychotherapeutic relationship.

The hypothesized relationship between positive attraction to therapy and movement toward self-descriptive independence over therapy is also supported. Those clients who are positively attracted see themselves as becoming more independent as therapy progresses. Of additional interest is the fact that a distinction can be made between self-descriptive and behavioral changes toward independence. While the attracted clients saw themselves as becoming more independent, this relationship was not found on the overt behavioral measure. Behaviorally, the attracted clients in the experimental group were still dependent at the time of the posttherapy testing, i.e., after 15 sessions of psychotherapy. It appears that attracted clients may be set to see themselves as changing, although their interpersonal interactions remain relatively constant. The motivation for this change in self-descriptive dependency may well be the desire to please the therapist and thus say what they think the therapist expects or would like them to say. Working with a sample of psychotherapists who received their training at the same university as the therapists of the present study, Peterson, Snyder, Guthrie, and Ray (1958) have demonstrated

TABLE 1
CORRELATIONS BETWEEN CLIENT PRETHERAPY
ATTRACTION AND DEPENDENCY

Pretherapy Attraction with:	Group		
	All Subjects	Experimental	Control
Pretherapy			
EPPS dependency	.501***	— ^a	—
situational dependency	.491***	—	—
Pre-post difference			
EPPS dependency	—	-.774***	-.508*
situational dependency	—	-.189	.002
Posttherapy			
EPPS dependency	—	.065	-.009
situational dependency	—	.542**	.241

^a Data obtained at this time preceded the assignment of subjects to experimental and control groups, hence pretherapy *r*'s were calculated across all 30 subjects.

* Significant at .06 level.

** Significant at .05 level.

*** Significant at .01 level.

that therapists of this training background show a great deal of attention to their clients when the latter's remarks demonstrate a preference to let others provide for the satisfaction of their needs. Should this differential between self-descriptive and overt behavioral test findings be corroborated in further research, serious doubt would be cast upon the common procedure of evaluating the effectiveness of psychotherapy by the *exclusive* use of self-descriptive measures.

Somewhat more puzzling is that the control clients, who received no formal psychotherapy, showed almost the same relation between attraction and self-descriptive movement toward independence. It should be noted that the control group *as a whole* showed no movement over therapy, either in the independent or the dependent direction. But still, when individual variation is considered, those in the control group who described themselves as becoming more independent over time were positively attracted toward therapy, while those who described themselves as becoming dependent tended to be those who were negatively attracted. The investigators can only speculate concerning the reasons for this relationship in the control group. Our present inclination is to view attracted clients as individuals who would interpret even minimal clinic contact (such as is involved in testing sessions) as benefiting them in some way.

TABLE 2

CORRELATIONS BETWEEN CLIENT ATTRACTION AND THERAPIST EXPECTANCY

Variables correlated	<i>r</i>
Preattraction and:	
TE ₅ ^a	.427
TE ₁₀ ^b	.144
TE ₁₅ ^c	.199
Difference-attraction and:	
TE ₅	.619*
TE ₁₀	-.137
TE ₁₅	.418
Postattraction and:	
TE ₅	.535*
TE ₁₀	-.162
TE ₁₅	.096

^a Fifth session therapist expectancy.^b Tenth session therapist expectancy.^c Fifteenth session therapist expectancy.

* Significant at .05 level.

A study by Barron and Leary (1955) appears to offer support for this contention. They state, with regard to wait-list control clients:

simply having committed oneself to participating in psychotherapy, and having had a reciprocal commitment from a clinic to afford psychotherapy, even though not immediately, represents a breaking of the neurotic circle. A force for change has already been introduced. In addition, the initial interview and the psychological testing may themselves be psychotherapeutic events, since during such sessions the patient makes some efforts to confront himself and his problems more objectively than he has in the past (p. 244).

A recent investigation by Goldstein (1960a) supports this finding.

Table 2 presents the correlations between therapist expectancy of client improvement, as obtained at five session intervals, and client pre- and posttherapy attraction, as well as the change in attraction over the course of therapy.

These findings indicate a significant relation between the expectation of client improvement held by the therapist early in therapy, and both the change in client attraction over the course of therapy and the magnitude of client attraction subsequent to the fifteenth session. None of the other correlations presented in Table 2 reached accepted levels of significance.

In addition to offering partial support for the hypothesis that therapist expectancy is a relationship maintaining aspect of the psychotherapeutic interaction, these findings raise the question as to why this should only be the case with regard to the therapist's early expectations (fifth session), and not his tenth and fifteenth session anticipations of client improvement. A study by Good (1952) furnishes a basis for differentiating "early" and "late" therapist expectations, a differentiation which appears to shed light on the present study's findings. He states:

Support was found for the hypothesis that in a relatively novel situation, generalization effects chiefly determine the expectancy held by *S*, and that as *S* has more experience with the specific task, expectancies develop which are a function of this task (p. 99).

In the present study, the expectations held by the therapist at the fifth session regarding client personality change may be more a function of his perceived success and failure with past clients than his feelings concerning his present client's progress. By the tenth session, however, their psychotherapeutic interaction is less "novel" and the major determiner of the therapists' expectations may have shifted from generalization effects to task effects. Kelly (1955), Lennard and Bernstein (1960), and Rotter (1954) have also noted significant temporal shifts in therapist expectancies over the course of psychotherapy.

The basis for the failure of late-therapy therapist expectations to be relationship maintaining would appear to be an important question for further research.

SUMMARY

The present investigation attempted to determine the extent to which client dependency and therapist expectation of client improvement can be considered relationship maintaining variables in the psychotherapeutic interaction. Thirty clients undergoing psychotherapy at a University Psychological Clinic were randomly assigned to "therapy" and "no-therapy" conditions and the 15 "therapy" clients were randomly assigned to 10 therapists. The therapists, for the most part, were advanced graduate students in clinical psychology at the Pennsylvania State

University. Testing, on measures developed or modified for the current study, took place pre- and posttherapy for all clients and after every five sessions for the therapists. Results of the study indicated a strong positive relation between client pretherapy attraction to the therapist and: (a) both client self-descriptive and behavioral dependency before therapy and, (b) client self-descriptive, but not behavioral, movement toward independence over the course of therapy. A similar but less marked relationship occurred in control group clients, offering further evidence for the therapeutic nature of such nonspecific clinic contacts as the intake interview and psychological testing.

An unexpected finding was the relatively high degree of relation between pretherapy attraction in therapy clients and their overt posttherapy dependency, a finding at variance with that obtained on self-descriptive instruments. In addition to other implications, this finding suggests caution in interpreting results in psychotherapy research which are based solely on one level of measurement.

Finally, partial support was obtained for the hypothesis that favorable therapist expectation of client improvement can function to maintain the therapeutic relationship.

REFERENCES

- BARRON, F., & LEARY, T. Changes in psychoneurotic patients with and without psychotherapy. *J. consult. Psychol.*, 1955, 19, 239-245.
- BERNARDIN, A. C., & JESSOR, R. A construct validation of the Edwards Personal Preference Schedule with respect to dependency. *J. consult. Psychol.*, 1957, 21, 63-67.
- BORDIN, E. S. Inside the therapeutic hour. In E. A. Rubenstein & M. B. Parloff (Eds.), *Research in psychotherapy*. Washington, D. C.: American Psychological Association, 1959. Pp. 235-246.
- BORGATTA, E. F. An analysis of three levels of response: An approach to some relationships among dimensions of personality. *Sociometry*, 1951, 14, 267-316.
- DOLLARD, J., & MILLER, N. E. *Personality and psychotherapy*. New York: McGraw-Hill, 1950.
- EDWARDS, A. L. *Edwards Personal Preference Schedule: Test and manual*. New York: Psychological Corporation, 1954.
- FITZGERALD, B. J. Some relationships among projective tests, interviews, and sociometric measures of dependent behavior. *J. abnorm. soc. Psychol.*, 1958, 56, 199-203.
- FORDYCE, W. E. Applications of a scale of dependency to concepts of self, ideal-self, mother, and father. Unpublished doctoral dissertation, University of Washington, 1953.
- GISVOLD, D. A validity study of the autonomy and deference subscales of the EPPS. *J. consult. Psychol.*, 1958, 22, 445-447.
- GOLDSTEIN, A. P. Patient's expectancies and non-specific therapy as a basis for (un)spontaneous remission. *J. clin. Psychol.*, 1960, in press. (a)
- GOLDSTEIN, A. P. Therapist and client expectations of personality change in psychotherapy. *J. counsel. Psychol.*, 1960, in press. (b)
- GOOD, R. A. The potentiality for changes of an expectancy as a function of the amount of experience. Unpublished doctoral dissertation, Ohio State University, 1952.
- HELLER, K. Dependency changes in psychotherapy as a function of the discrepancy between conscious self-description and projective test performance. Unpublished doctoral dissertation, Pennsylvania State University, 1959.
- KELLEY, H. H. The effects of expectations upon first impressions of persons. *Amer. Psychologist*, 1949, 4, 252. (Abstract)
- KELLY, G. A. *The psychology of personal constructs*. New York: Norton, 1955.
- LENNARD, H. L., & BERNSTEIN, A. *The anatomy of psychotherapy*. New York: Columbia Univer. Press, 1960.
- LIBO, L. *Picture impressions: A projective technique for investigating the patient-therapist relationship*. (Department of Psychiatry Publication Series) Baltimore: Univer. Maryland Medical School, 1956.
- LIBO, L. The projective expression of patient-therapist attraction. *J. clin. Psychol.*, 1957, 13, 33-36.
- MURRAY, H. A. *Thematic Apperception Test: Pictures and manual*. Cambridge: Harvard Univer. Press, 1943.
- PETERSON, A., SNYDER, W. U., GUTHRIE, G. M., & RAY, W. S. Therapeutic factors: An exploratory investigation of therapeutic biases. *J. counsel. Psychol.*, 1958, 5, 169-173.
- ROGERS, C. R., & DYMOND, R. G. *Psychotherapy and personality change*. Chicago: Univer. Chicago Press, 1954.
- ROSENTHAL, R. Research in experimenter bias. Paper read at American Psychological Association, Cincinnati, September 1959.
- ROSENZWEIG, S. Rosenzweig Picture-Frustration Study, revised form for adults: Test and manual. St. Louis: Author, 1947.
- ROTTER, J. R. *Social learning and clinical psychology*. New York: Prentice-Hall, 1954.
- UHLHUTH, E. H., CANTER, A., NEUSTADT, J. O., & PAYSON, H. E. The symptomatic relief of anxiety with meprobamate, phenobarbital and placebo. *Amer. J. Psychiat.*, 1959, 115, 905-910.
- ZUCKERMAN, M., & GROSZ, H. J. Suggestibility and dependency. *J. consult. Psychol.*, 1958, 22, 328.

(Received July 26, 1960)

THE EFFECTS OF PSYCHOTHERAPY ON SELF-CONSISTENCY:

A REPLICATION AND EXTENSION¹

ROSALIND DYMOND CARTWRIGHT²

University of Colorado

In a previous paper (Cartwright, 1957), it was proposed that one of the common themes of psychotherapy with psychoneurotic patients is a search for a stable identity. In that paper the author put the problem in these terms:

If selves or roles are thought of as characteristics which are particular to specific interactions or classes of interaction, then the area of self is that core of characteristics which is common over *N* situations. In these terms the pre-therapy client can be thought of as one whose self is very small. He seems to have diverse selves in relation to others with whom he interacts, but the common core, the essential me is so restricted in scope as to leave him puzzled by the question "Who am I?"

It was hypothesized that the self the pre-therapy client describes would differ considerably depending on his interactional referent, while the posttherapy client's self-description would have more stability in terms of a larger core of characteristics which remain consistent in emphasis, regardless of the particular other involved. Of course it is most likely that the healthy individual will retain some variability, that there is an optimal point here short of rigidity.

Specifically, the 1957 study hypothesized that, if asked to select three people of major importance to them and to describe themselves on a *Q* sort as they are with each of these people in turn, a group of clients would show more variability among these sortings

before therapy began than after therapy had been completed; successful clients would change more than failure clients, and clients before therapy would have more variability in their self-descriptions than a group of control subjects, but would not differ from controls after therapy had been concluded.

Although the 1957 study supported the hypothesis outlined above, there were several reasons why it was felt that the study should be replicated and extended. (a) The sample was a very small one: 10 experimental subjects (5 successful and 5 unsuccessful cases) and 10 controls. (b) The controls were only tested at one point in time, as it was assumed that they would not change in self-consistency. This should be experimentally established. (c) The relations among these various self-to-other sortings were not tested. For these reasons the present study was undertaken.

METHOD

Sample

The present sample consists of 19 experimental subjects³ and 20 controls. Originally 30 subjects who applied to the University of Chicago Counseling Center for psychotherapy were asked to participate in this research study. Of these, eight discontinued therapy before completing the minimum number of interviews, which was set at six. An additional three cases of those completing therapy were lost by failure to make contact with them for posttherapy testing.

The control sample was selected to match the experimental sample in age, sex distribution, and student and nonstudent status. As they were to be controls for the variable therapy, they were selected as never having had, not now having, and having

¹ The data for this study were collected under a research grant from the Ford Foundation to the University of Chicago Counseling Center.

² The author wishes to thank John L. Vogel, now of the Department of Psychiatry, University of Washington, Seattle, who did all of the testing of the subjects.

³ This sample is identical to that reported on in the paper by Cartwright and Vogel (1960).

no immediate intention of having psychotherapy. Table 1 compares the samples in the two studies.

Procedure

The procedure differed slightly from that used in the previous study in that the subjects were first asked to complete a "plain" self-sort. The 100 item Butler and Haigh Q cards (1954) were again used as the primary instrument. Each subject was first instructed to sort the cards to describe himself as he is today from those items which are most like him to those which are least like him according to the required 1, 4, 11, 21, 26, 21, 11, 4, 1 distribution. Next a TAT was administered. Following this the procedure was identical to that previously reported. The instructions to the subjects were:

I would like you to think of three people who are very important to you. They can be of any age or relation to you like father, mother, child, friend, or boss. They can be people you like or dislike, or some of each. You don't have to pick on any basis except they be people who are very important to you and with whom you have real interaction.

The subject was then given a code sheet on which to enter code names for his choices opposite their role relation to him. He was then given a deck of Butler and Haigh Q cards and instructed to sort them to describe himself as he is in his relationship to the first person on his list.

Sort these cards to describe yourself as you see yourself to be in your relationship to _____, from those cards which are least like you as you are with him (her) to those that are most like you in your relations with him (her). Keep this question in mind while you are working: If I were only the person I am with _____, what would I be like?

Following this sorting, the subject was given another set of the same cards and asked to sort again to describe himself as he is with the second person on his list. Then this was repeated a third time for the third person of major importance to him. All in all then, there were four sortings of the

TABLE 1

COMPARISON OF THE EXPERIMENTAL AND CONTROL SAMPLES IN THE 1957 AND 1959 STUDY

Sample	Sex			Age		Number of interviews	
	N	M	F	Mean	Range	Mean	Range
E 1957	10	7	3	25	18-42	26	8-65
C 1957	10	7	3	26	19-44		
E 1959	19	9	10	27.6	18-41	36.6	8-97
C 1959	20	10	10	29.3	19-53		

TABLE 2

COMPARISON OF THE MEAN ITEM VARIANCES FOR THE 1957 AND 1959 STUDIES

Sample	Pretherapy	Posttherapy
E 1957	.969	.766**
E 1959	.939	.763*
C 1957	.736	Not tested
C 1959	.763	.566****
S 1957	.849	.546***
S 1959	.940	.580***
F 1957	1.088	.936
F 1959	.938	.928

* Significant at .05 level.

** Significant at .02 level.

*** Significant at .01 level.

**** Significant at .001 level.

same Q items using the same distribution: one self-sort and three of his conception of himself in three different relationships.

For the experimental subjects these procedures were repeated when they had completed their therapy. For 10 of the control subjects they were repeated after a period of 6 months (to match the mean of 18 weeks of therapy).

ANALYSIS AND RESULTS

To test the major hypothesis of increased self-consistency following psychotherapy, the same method reported previously was employed. Briefly the mean item variance over the three self-in-relationship sortings was computed by a method developed by Cartwright (1956a). The hypothesized differences between these means were then tested using *t*. The *t* between pre- and posttherapy mean for the experimental (E) group was 2.100, significant at the .05 level.⁴

The *t* between the first and second testing of the control (C) group was 4.394, significant at the .001 level.

From these results it would appear that a second testing of either experimental (therapy) or control (no therapy) subjects will show a significant increase in consistency of sorting.

In the 1957 study, five successful therapy cases were found to decrease their mean item variance significantly while five failure cases did not. Using the same criterion of successful therapy for the present sample

⁴ All significance levels are for two-tailed tests unless otherwise stated.

TABLE 3

MEAN *Q* ADJUSTMENT SCORES OF SELF-SORT OF CONTROL, SUCCESS, AND FAILURE GROUPS

Group	<i>Q</i> Adjustment score		Direction of change in adjustment and self-consistency	
	Test 1	Test 2	Same	Different
Control	49.3	50.5	5	5
Success	36.7	49.4****	9	0
Failure	36.1	38.3	6	4

**** Significant at .001 level.

yielded only 3 cases of success out of the 19. This criterion (based on therapists' ratings of adjustment, change in adjustment, and success of the therapy) was, therefore, reluctantly abandoned for one that would split the group more equally. The new criterion was based on change in test performance from pre- and posttherapy time. The 19 experimental subjects were ranked on the extent of their improvement on two test scores: the *Q* Adjustment score (Dymond, 1954a) and a mental health rating of the TAT (Dymond, 1954b). The 9 subjects who improved most on this combined ranking of change were called the success (S) group and the 10 who ranked lowest were called failures (F). Table 2 compares the mean item variances for the groups and subgroups involved in the original and in this replication study. It is clear from Table 2 that failures as a group do not increase in self-consistency on a second testing.

While the extent of the replication of the mean scores in the two studies is striking, particularly since such small groups were involved, the meaning of the change for the therapy subjects is obscured by the highly significant change in the group which was not in treatment. Testing for the significance of the difference between the changes in the S group and the C group yielded a nonsignificant *t* (1.644). However, inspection of the array of change scores showed that the range of changes was considerably greater for the S group than for the Cs. An *F* ratio was computed and found to be 3.30, significant at the .05 level. A comparison of the disper-

sion of the changes between the total experimental group and the C group showed the E group's dispersion to be much wider, of course. The *F* ratio here was 13.19, significant beyond the .001 level. These findings of the significant difference in the variances of changes between E and C and between S subgroup and C is similar to the finding reported by Cartwright (1956b) on his re-analysis of the data reported by Barron and Leary (1955). He concluded that clearly there was more *change* with regular formally defined therapy than without it, but in both directions. Similarly, in this study it can be argued that although both E and C groups changed significantly in the direction of increased self-consistency over similar time periods, those in therapy showed wider changes. It seems that therapy has more impact to change persons on this measure than the unknown influences which brought about the change in the no-therapy group.

Further light is shed on the meaning of these changes by looking at the *Q* Adjustment scores of the plain self-sort. If the change toward greater consistency is to be evaluated positively, the more stable self should be one the subject can live with more happily. Therefore it should be accompanied by a higher Adjustment score of the self-description. This is not a necessary relationship. It is possible for the self-in-relationship to other sortings to become less varied without an accompanying change towards better adjustment. This is what appears to happen in the control cases.

Several important points emerge from an inspection of Table 3.

1. The C group does not improve in adjustment despite the fact that it does increase in self-consistency. The improvement in the Adjustment score of the E group as a whole is significant beyond the .01 level; in the S subgroup at the .001 level.

2. In only 5 of the 10 C cases do the two measures move in the same direction, whereas the ratio is 15 of 19 for the E group, and 9 out of 9 for the S subgroup. It seems likely then that the increase in self-consistency for the C cases must be interpreted differently, perhaps due to less motivation for making fine distinctions on the second test. The in-

structions certainly set the subjects to look for differences in their ways of viewing themselves in their interactions. If the subject is largely consistent in his self-picture over various interactional settings, the instructions might cue him to make small distinctions which are relatively unimportant to him, and so not maintained on Test 2. If this explanation is tenable, it would be expected that the C group would differ from the S group in various ways. (a) They would have fewer items with large discrepancies on Test 1 than S subjects, and more small discrepancy items. (b) They would show less proportional change in their large discrepancy items than the S group. (c) The increased consistency of the C subjects would not represent any major change in self-definition. Items which are consistent on Test 2 which were not previously consistent will have changed less in their position of importance to the C subjects than to the S cases.

To test these notions, the extent of the discrepancy over the three sortings for each item was tabulated. Items with a discrepancy of 0-2 Q points were called Small Difference items. Items with discrepancies of 3 or more Q points were called Big Difference items. As these predictions were all directional in nature, one-tailed tests were used. All were confirmed at the .05 level or better.

Looking at the new consistent items for the two groups, it was found that the S group drew their new self items from a different source than did the C group. If items sorted at the extreme scale positions (0,1,2 = least like me; and 6,7,8 = most like me) are assumed to be more important to the person than those sorted at the middle of the forced normal distribution, an item which was not previously an important self item (extreme in placement and consistent in position) could have been previously extreme but inconsistent or previously central. Table 4 shows the previous position of the new extreme consistent items for the S and C groups.

Without doubt the new stable elements of the self which are of importance to the successful therapy people are significantly more often items which have changed with respect to their importance. It is more often the case

TABLE 4
PREVIOUS POSITION OF NEW EXTREME
CONSISTENT ITEMS

Success group		Control group	
Central	Extreme	Central	Extreme
11	0	3	9
9	3	2	6
14	5	2	7
12	3	5	4
8	2	2	6
4	2	0	2
8	6	2	6
8	1	3	10
7	2	5	5
Mean		Mean	
9.0	2.6	2.7	5.9

Note.—Success: Central-Extreme $t = 5.161^{***}$
 Control: Central-Extreme $t = 3.298^{***}$
 Central: Success-Control $t = 6.262^{***}$
 Extreme: Success-Control $t = 3.367^{***}$

*** Significant at .01 level or beyond.

that these items have gained an importance over the therapy period, rather than that they were important previously and have only gained in stability or consistency. For the controls, the reverse is true. The new stable items at the extremes of the sorts on Test 2 were significantly more often also at the extremes on Test 1, although at that time inconsistently sorted. Here, then, is another indication that the increased consistency of these two groups represents different processes. The successful therapy cases have changed their definitions of themselves in their important relationships, while the controls seem only to have consolidated the definitions previously made.

If this proposed explanation holds, that the increased consistency of the two groups is essentially different, that therapy has an impact which shifts the meaning of the self as well as exercising a consolidating influence, then the particular selves that the S group describes should reflect the shift in meaning in their Adjustment scores in their relation to others. If the C group's consistency is due only to a lack of motivation for making fine distinctions, then no essential change of the adjustment of self in relation to others would be expected. Since each subject was free to choose the relationships

TABLE 5
MEAN ADJUSTMENT SCORES FOR SELF-IN RELATIONSHIP TO MOTHER, FATHER, AND
OTHERS OF SUCCESS, FAILURE, AND CONTROL GROUPS

Group	Mother		Father		Other	
	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2
Success	38.3	49.8	46.0	53.2	44.4	52.1
Failure	32.2	37.2	39.4	36.2	43.3	34.8
Control	52.7	54.8	42.3	45.5	46.9	47.9

Note.—Success: Mother Test 1–Mother Test 2 $t = 4.063^{***}$
 Others Test 1–Others Test 2 $t = 3.660^{***}$
 Failure: Others Test 1–Others Test 2 $t = 2.260^*$
 Mother Test 1–Others Test 1 $t = 2.206^*$
 Control: Mother Test 1–Father Test 1 $t = 2.369^*$
 Mother Test 2–Father Test 2 $t = 1.917$
 Mother: Success Test 1–Control Test 1 $t = 3.237^{***}$
 Failure Test 1–Control Test 1 $t = 4.053^{***}$
 Success Test 2–Failure Test 2 $t = 1.930$
 Failure Test 2–Control Test 2 $t = 3.154^{***}$
 Father: Success Test 2–Failure Test 2 $t = 3.157^{***}$
 Others: Failure Test 2–Control Test 2 $t = 2.689^{**}$
 Success Test 2–Failure Test 2 $t = 2.409^*$

* Significant at .05 level.

** Significant at .02 level.

*** Significant at .01 level.

of importance to him, it is hard to make strict comparisons. For this reason relationships were only categorized as father, mother, and others. Table 5 gives the mean Adjustment scores for these categories at the two testing points.

Some of the significant relations were: (a) S cases improve in the adjustment of the relationship to their mothers between Test 1 and Test 2. Neither F cases nor C cases show any significant change. (b) Both E groups are more poorly adjusted in their relationship to their mothers on Test 1 than C cases. S cases are not significantly different after therapy from the adjustment level of the C group on their second test. S cases border on being significantly higher in adjustment on their second test than F cases who are still significantly poorer in mean adjustment than Cs. (c) There is no significant difference between the adjustment level of self-in-relationship to mother and father for either E group at either time. However, the self-in-relationship to father of the C group is significantly lower than their self-in-relationship to mother on Test 1, and this borders on being a significant relationship on Test 2 as well. (d) None of the groups changes significantly in the adjustment of the self-in-relationship to father. The F group is lowest of the three in mean ad-

justment on Test 1, and this drops on Test 2 to be significantly lower than the S group. (e) S cases improve their mean adjustment score in relation to others significantly over the therapy time, whereas F cases change significantly towards poorer adjustment. On Test 2 Fs are significantly lower in adjustment to others than both Ss and Cs.

It would appear from these results that patients, as distinct from controls, come to therapy with disturbance in their self-to-mother relationship. Patients who subsequently do not succeed in therapy start out with poor adjustment of their self-image-in-relation to both parents while those who succeed start with strength from their self-to-father image.

Since it seemed likely that some interesting conclusions could be drawn from these results concerning the kinds of problem relationship that persons bring to therapy and those which do and do not get resolved with therapy of this kind, it seemed important to check these data as far as possible with those from the previous study. This is difficult since a different criterion was employed and the total number of cases was only 10. For this latter reason, the comparison was made for the total E group.

The first study shows some of the same patterns as found in the present sample:

TABLE 6

COMPARISON OF MEAN ADJUSTMENT SCORES FOR 1957
AND 1959 STUDIES

Sample	Mother		Father		Other	
	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2
E 1957	31.4	42.8	30.8	41.0	37.0	43.7
E 1959	35.0	43.0	43.4	46.0	43.8	43.0
C 1957	46.6		38.0		47.2	
C 1959	52.7	54.8	42.3	45.5	46.6	47.2

justment of the self-in-relation to their mothers, and this is not true for either sex group of control subjects. Also, it is obvious that of the clients, it is the males that have father problems that bring them to therapy. Further, therapy of this kind improves the adjustment of the self-in-relationship to the mother for both sexes, but not significantly for males, and there is little change in the father relationship for either sex.

SUMMARY AND DISCUSSION

This paper has reported a study designed to replicate and extend a previous study of the effects of psychotherapy on self-consistency. In both experiments the samples consisted of psychoneurotic subjects who had applied to the University of Chicago Counseling Center for client centered psychotherapy. Both studies employed matched control subjects who were not motivated for therapy. The original study found the therapy group to have lower self-consistency than controls before therapy, and to have increased their consistency at the completion of therapy to the level of the control group.

The replication study confirmed the findings of the first, but found that controls also increase their self-consistency on Test 2 significantly over their Test 1 level. An analysis of the kinds of change involved showed different processes at work. The increased consistency of the controls did not represent much redefinition of the self-in-relationship

(a) The mean adjustment of the C group's self-in-relation to father is again lower than their mean adjustment of self-in-relation to mother, although this does not reach statistical significance. (b) Patients in the 1957 study were also lower in mean adjustment of self-in-relation to mother than were controls. Speaking loosely then, both studies show that clients come to therapy with mother problems which are not present in control subjects. Control subjects may have father problems, but these do not seem to be sufficiently disturbing to bring them to therapy. The good adjustment of the self-in-relation to the mother seems to be the distinguishing mark between patients and those who are not patients. However, it was thought that perhaps the sex of the subjects might be an important variable in these relationships. Table 7 gives the mean adjustment scores for the two sexes in the current study.

Obviously from Table 7 it is clear that clients of both sexes have problems of ad-

TABLE 7

COMPARISON OF MEAN ADJUSTMENT SCORES OF MALES AND FEMALES

Sample	Mother		Father		Other	
	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2
E 1959						
Males (<i>N</i> = 9)	31.8	39.8	38.7	41.4	42.5	41.2
Females (<i>N</i> = 10)	37.8	45.8	50.0	52.8	45.9	45.1
C 1959 ^a						
Males (<i>N</i> = 10)	50.1		44.9		50.0	
Females (<i>N</i> = 10)	51.3		44.0		44.6	

Note.—Males: Mother Test 1-Test 2 $t = 1.473$
Females: Mother Test 1-Test 2 $t = 4.997^{***}$
^a All 20 cases that were given Test 1 are reported here to maximize the *N*.
^{***} Significant at .01 level.

to the important others but rather a consolidation of the definitions previously made. In addition, there was no accompanying change in the adjustment level of the self for the control group. In the successful therapy group the new self-consistent items showed significantly more shift in importance in the various relationships so that it could be stated that redefinition was taking place. Moreover, the overall adjustment level of the self-sort improved significantly.

Looking at the particular interactional referents, it was found that clients were more poorly adjusted in their self-to-mother image than controls, and this relationship held for both studies and for the experimental group when stratified both by outcome of therapy (success and failure), and by sex (males and females). This appears to be a reliable distinguishing characteristic of subjects who apply for psychotherapy (at least of this kind), and who stick with it beyond six interviews.

Control cases in the present study were significantly poorer in adjustment of their self-image in relation to their fathers than to their mothers, and this bordered on being a significant difference for the original sample of controls as well. It seems likely, then, that it may be "normal" to have a more poorly adjusted image of oneself in relation to father than to mother, but providing the self-to-mother image is sufficiently well adjusted, there is no internal pressure to seek psychotherapy. This finding held equally for both sexes of control subjects in the present study. Unfortunately, the *N* in the previous study was too small to test it for that group.

In the present study, males were found to enter therapy with significantly poorer adjustment of self-in-relationship to father than did females. Successful cases improved the adjustment of their self-sort in relation to their mother, but there was no improvement of adjustment of the self-sort in relation to the father for any group in this study. This type of therapy, then, seems to be effective in resolving mother problems, but does not appear to change the self-to-father image.

Perhaps clients with mother problems self-select client centered therapy as meeting their needs of working out their problems of adjustment to a maternal figure. The therapist in this type of therapy is supposed to be warm, accepting, empathic, understanding, nondirective and to stress the feeling aspects of the communications: to play, essentially, a very feminine role. Or perhaps most psychoneurotic outpatients would be found to have disturbance in the mother area and some, namely, males, in the father area as well. But those who are in a more directive type of therapy might be found to have a reversed pattern of changes: improvement in the father area and little change in the mother area. These questions remain open for further research evidence.

REFERENCES

- BARRON, F., & LEARY, T. F. Changes in psychoneurotic patients with and without psychotherapy. *J. consult. Psychol.*, 1955, 19, 239-245.
- BUTLER, J. M., & HAIGH, G. V. Changes in the relation of self-concepts and ideal concepts. In C. R. Rogers & Rosalind F. Dymond (Eds.), *Psychotherapy and personality change*. Chicago: Univer. Chicago Press, 1954. Pp. 55-75.
- CARTWRIGHT, D. S. A coefficient of consistency over T Q sorts. *Counseling Center discussion paper*, 1956, 2, No. 14. (a)
- CARTWRIGHT, D. S. Note on "Changes in psychoneurotic patients with and without psychotherapy." *J. consult. Psychol.*, 1956, 20, 403-404. (b)
- CARTWRIGHT, ROSALIND D. Effects of psychotherapy on self-consistency. *J. counsel. Psychol.*, 1957, 4, 15-22.
- CARTWRIGHT, ROSALIND D., & VOGEL, J. L. A comparison of changes in psychoneurotic patients during matched periods of therapy and no therapy. *J. consult. Psychol.*, 1960, 24, 121-127.
- DYMOND, ROSALIND. Adjustment changes over therapy from self-sorts. In C. R. Rogers & Rosalind F. Dymond (Eds.), *Psychotherapy and personality change*. Chicago: Univer. Chicago Press, 1954. Pp. 76-84. (a)
- DYMOND, ROSALIND. Adjustment changes over therapy from Thematic Apperception Test ratings. In C. R. Rogers & Rosalind F. Dymond (Eds.), *Psychotherapy and personality change*. Chicago: Univer. Chicago Press, 1954. Pp. 109-120. (b)

(Received July 26, 1960)

SEXUAL SYMBOLIC RESPONSE IN PREPUBESCENT AND PUBESCENT CHILDREN

AUSTIN JONES¹

University of Pittsburgh

The Freudian hypothesis that pointed, elongated objects are symbolic of the penis and that rounded or enclosing objects are symbolic of the vagina has received inferential experimental support in a recent study (Jones, 1956). Adult subjects were asked, essentially, to ascribe masculine or feminine personality to each of a series of simple geometric figures. The figures were of two classes: pointed or elongated, and rounded or enclosing. Subjects responded in a manner consistent with the Freudian hypothesis significantly more often ($p < .001$) than would be expected by chance alone. Male subjects were found to respond significantly more consistently with the hypothesis than females. In discussing the latter finding, it was noted that available evidence (Kinsey, Pomeroy, Martin, & Gebhard, 1953) suggests that the sexual response of men is more readily conditioned to a wide variety of visual stimuli than is that of women. The symbols employed in the study were modifications of those used earlier by Levy (1954), who failed to find support for the Freudian hypothesis in an experiment involving the matching of symbols and male and female given names, and the paired associate learning of like-sex and unlike-sex pairs of symbols and names. The subjects of that study were fifth grade children, most of whom were presumably prepubescent. In light of the subsequent positive findings with adult subjects, it appears possible that age differences in the sexual symbolic response might account partly for the negative results of the earlier study. Developmental studies of sexual symbolism of an experimental nature are apparently totally lacking.

¹ The author wishes to thank William Bendig, Ruth Goodman, and Melvin Manis for helpful comments and advice.

The purpose of the present study was to assess the strength of the sexual symbolic response in children of various ages up to the time, roughly, of pubescence. The attitude of the study was exploratory; it is possible to formulate rather contradictory hypotheses as to the function relating age and strength of response. One view would hold that, symbolic behavior being generally a matter of socially acquired learning, the strength of the sexual symbolic response ought to increase with increasing age and increasing exposure to the socially determined symbols. Also, symbolic behavior generally may be conceived of as evidence of socially imposed inhibitions of more direct expression; the young child, having lived fewer years under the inhibiting influence of civilization, would be expected to show less use of symbolism and relatively greater directness of sexual expression than would the older child or adult. A different hypothesis emerges when we consider the data regarding drive level and generalization phenomena. As drive increases, the height of the generalization gradient is raised and discrimination correspondingly reduced. Beach (1942), for example, increased the sex drive of a group of male rats by injections of testosterone, with the result that the animals made sexual responses to a markedly increased range of stimuli—thus failing to discriminate what function “normally” as sexual and non-sexual stimuli. Similarly, we would expect that children approaching pubescence, i.e., experiencing an increase in sex drive, would demonstrate a decrement in discrimination between male and female sexual cues. Their responses to symbols would be less consistent with the Freudian formulation than those of children somewhat younger. The following

experiment was designed to clarify the issues involved in these contrasting hypotheses.

METHOD

Subjects

Since the intent of the experiment was to study the sexual symbolic response as a function of degree of sexual maturation, a strictly chronological arrangement of subject groups which includes both sexes is less meaningful than a categorization which takes into account the different maturation schedules of males and females. Due to the onset of puberty approximately 2 years earlier in females than in males, roughly equivalent levels of sexual maturation are obtained by grouping the males of a given age with females 2 years younger. Four such "sexual maturation" groups were constituted using boys and girls in a public school system. The "youngest" sexual maturation group consisted of second grade girls and fourth grade boys; the next two groups, fourth grade girls plus sixth grade boys, and sixth grade girls plus eighth grade boys. The final group was planned to be eighth grade girls and tenth grade boys; but administrative permission could not be obtained for the utilization of high school (and hence tenth grade) students. Ninth grade boys were available, however, so that the final sexual maturation group was approximated by pairing eighth grade girls and ninth grade boys. Ten boys and 10 girls were selected randomly at each of the four levels, making a total of 80 subjects. The mean ages of the five grades from which subjects were drawn (grades two, four, six, eight, nine) were 8.5, 10.5, 12.3, 14.4, and 15.5, respectively. Thus the subjects of the first sexual maturation level are approximately four years pre-pubescent, while at the fourth level most subjects of both sexes have attained puberty.

Materials

The stimuli in this experiment were the 10 sexual symbolic figures adapted by Jones (1956) from those used initially by Levy (1954). Five are pointed or elongated ("male") and five are rounded or enclosing ("female"). The figures are essentially "abstract" or without specific object identity; they are variations of circles, pyramids, rods, cubes, etc. The figures are reproduced in black ink on individual white cards. In the prior study supporting the Freudian view with adult subjects, one of the presumed female symbols was shown not to be so. The figure was included in the present administration in order to make the procedures of the two studies comparable for clarity of discussion, although responses to the nonfunctional figure did not enter into the score variable.

Procedure

The subjects were seen individually by the author. The experimental procedure was second in a series of other experimental procedures, allowing time for

the establishment of optimal rapport. Subjects were presented the following instructions orally. They are the instructions employed in the prior study with adult subjects except for editing to lower the vocabulary level.

You know, lots of times *things* kind of seem like *people* to us—kind of remind us of people somehow. Many things around the house, for instance, or almost anything you know real well. Some things seem more like *men* and some things seem more like *women*. I have a set of cards here which have designs or pictures on them. I'm going to show them to you one at a time—and you tell me which they remind you of more—men or women. Just use your imagination and tell me the first answer you think of. Don't stop to think about it; just tell me the first thing you think of.

Following the instructions, the experimenter presented the 10 figures in an order randomized for each subject. The subjects were required to respond within approximately 2 seconds.

RESULTS

The data of the experiment are presented graphically in Figure 1. The percentage of responses plotted there is the percentage (out of five male symbols or out of four female symbols) which were consistent with the Freudian hypothesis. The data are plotted separately for male and female subjects and for their response to the two types of symbols, male and female. Data for the normal adults of the prior study (Jones, 1956) are plotted

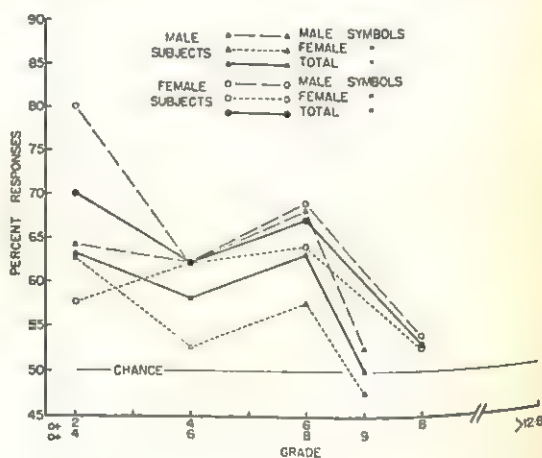


FIG. 1. Mean percentage of responses consistent with the Freudian hypothesis as a function of maturation level. (Indicated by the grade scale adjusted so as to equate maturation level for the two sexes. The two points plotted at the extreme right represent the means for the college subjects of the prior study.)

for comparison. Table 1 summarizes an analysis of variance of trends (Grant, 1956) over the four sexual maturation levels. The arc-sine transformation for proportions was employed (Walker & Lev, 1953). Differences between maturation levels (linear component) attained the .06 level of significance, and differences between symbol types the .01 level. Differences between subject sex groups failed to reach significance, as was true of the various interactions.

An additional analysis of variance was performed which compared the pooled child data of the present study with the adult data of the prior one. This analysis, summarized in Table 2, was thus based upon two age classes and, as before, two sex classes and two symbol type classes. Scores again were the arc-sine transformations of the proportions of "correct" responses. Significant differences were demonstrated between the two age groups and between symbol types ($p < .01$ in each case). The Age \times Sex interaction approached conventional significance levels ($p < .08$). All other interactions were nonsignificant.

TABLE 1

ANALYSIS OF VARIANCE OF PROPORTIONS OF "CORRECT" RESPONSES FOR EACH SYMBOL TYPE

Source	df	MS	F
Between maturation levels	3	7,704	2.18
linear	1	13,489	3.82*
quadratic	1	1,967	
cubic	1	7,657	2.17
Between sexes	1	3,600	1.02
Sex \times maturation level	3	59	
Between subjects within groups	72	3,532	
Between symbol types	1	8,717	7.80***
Symbol type \times maturation level	3	744	
Symbol type \times sex	1	21	
Symbol type \times maturation level \times sex	3	2,226	1.99
linear	1	3,024	2.70
quadratic	1	3,288	2.94
cubic	1		
Pooled symbol type \times subjects within groups	72	1,118	

Note.—Data in the form of arc-sine transformation for proportions.
 * $p < .06$.
 *** $p < .01$.

TABLE 2

ANALYSIS OF VARIANCE COMPARING THE CHILD DATA OF THE PRESENT STUDY WITH THE ADULT DATA OF THE PREVIOUS STUDY

Source	df	MS	F
Between ages	1	72,904	21.44***
Between sexes	1	320	
Age \times sex	1	10,682	3.14**
Between subjects within groups	96	3,400	
Between symbol types	1	6,083	4.90***
Symbol type \times groups	3	3,271	2.63**
Pooled symbol type \times subjects within groups	96	1,241	

Note.—Data in the form of arc-sine transformation of proportions of "correct" responses for each symbol type.

** $p > .05$.

*** $p < .01$.

DISCUSSION

The finding of principle interest has to do with the differences between maturation levels, the linear component of which was significant at the .06 level. An appropriate procedure, when findings fall very slightly short of the accepted .05 level, would be to replicate the experiment within the same population so as to provide a more definite evaluation when it is reported. Such a replication would have been desirable here but was not possible for administrative and public relations reasons. Consequently, the finding will be discussed tentatively here as significant, with the obvious comment that additional assessments of the relationship are needed.

The overall downward trend in sexual symbolic response over the four maturation levels of the present study supports the hypothesis that increased sexual drive as puberty approaches is accompanied by a raising of generalization gradients and an attendant decrement in discrimination between symbols of "maleness" and "femaleness." The four maturation levels extend, roughly, from 4 years prior to average age of pubescence, to 1 or 2 years beyond it. Sexual symbolic response during that period decreased from 66% to 52% (approximately chance expectancy). This trend was a highly consistent one, occurring in the responses of both male

and female subjects to both male and female symbols. Although it is not possible to reject entirely the hypothesis that socially acquired learnings tend gradually to instill the sexual symbolic response, it appears that their effect, if any, is obscured in the years approaching puberty by the relationship between increased drive and lowered discrimination. (The curious upward trend in the data of the third maturation level [approximately age 12 for girls, 14 for boys; see Figure 1] fails to reach conventional significance levels as tested by the cubic component of variance associated with maturation levels.)

The finding of significant differences between symbol types, with male symbols being responded to "correctly" more often than female symbols, confirms a trend in the data of the previous study (Jones, 1956). Whether or not this is a phenomenon of some generality remains unclear, since no attempt was made to equate the particular male and female symbols used for the degree of stimulus generalization which they represented. It is entirely conceivable that another set of alleged sex symbols might lead to a contrary finding. What is suggested, possibly, is that it is easier to draw a male sex symbol than a female sex symbol.

Comparison of the child data of the present study with the adult data of the previous one (Table 2) showed a significant increase in sexual symbolic response at young adult years over the highest frequency obtained in any of the prepubescent years. The frequency of response was 82% for the adult group (mostly in their early twenties) as compared with a high of 66% for children approximately 4 years prepubescent. Although sex drive is generally believed to decrease slightly between adolescence and the early twenties, it does not seem plausible to regard the associated lowering of generalization gradients as the sole cause of the striking increase in symbolic response. A more plausible but clearly conjectural explanation would be that following puberty individuals undergo a period of socially enforced discrimination training with respect to sexual cues and symbols, a degree of discrimination training not present in the prepubescent years. Thus, the overall develop-

mental history of the symbolic response, it is suggested, involves two phases with, most probably, different determinants. For at least 4 years prior to puberty, there is a continuous decrease in sexual symbolic response that is interpreted as a result of decreased discrimination attendant upon increased drive. By the early twenties, however, the frequency of sexual symbolic response has exceeded its prepubescent maximum, presumably as a function of the increased discrimination training implicit in the highly focused social control of adolescent heterosexual behavior. The verification of such an explanatory principle rests, of course, upon empirical research yet to be carried out.

SUMMARY

The purpose of the present study was to assess the strength of the sexual symbolic response in children of various ages through pubescence. Sexual symbolic response was defined as the designation of pointed, elongated figures as male and of rounded, enclosing figures as female. The figures employed were relatively abstract geometric forms taken from earlier studies of Jones and of Levy.

The mean frequency of sexual symbolic response was 66% for children approximately 4 years prepubescent. By the age of puberty or a year or 2 beyond, however, response had dropped sharply to about chance expectancy. The finding was interpreted as an effect of decreased discrimination associated with heightened sexual drive as puberty is approached. The data were compared with the performance of adults on the same task reported in a prior study. By the age of about 22, the frequency of the sexual symbolic response (82%) was shown to climb beyond the prepubescent maximum, perhaps as a function of the increased discrimination training implicit in the highly focused social control of adolescent heterosexual behavior.

REFERENCES

- BEACH, F. A. Effects of testosterone propionate on copulatory behavior of sexually inexperienced male rats. *J. comp. Psychol.*, 1942, 33, 227-247.

- GRANT, D. A. Analysis-of-variance tests in the analysis and comparison of curves. *Psychol. Bull.*, 1956, 53, 141-154.
- JONES, A. Sexual symbolism and the variables of sex and personality integration. *J. abnorm. soc. Psychol.*, 1956, 53, 187-190.
- KINSEY, A. C., POMEROY, W. B., MARTIN, C. E., & GEBHARD, P. H. *Sexual behavior in the human female*. Philadelphia: Saunders, 1953.
- LEVY, L. H. Sexual symbolism: A validity study. *J. consult. Psychol.*, 1954, 18, 43-46.
- WALKER, HELEN, & LEV, J. *Statistical inference*. New York: Holt, 1953.

(Received July 26, 1960)

RELATIONSHIPS BETWEEN 1960 STANFORD-BINET, 1937 STANFORD-BINET, WISC, RAVEN, AND DRAW-A-MAN

BETSY WORTH ESTES, MARY ELLEN CURTIN, ROBERT A. DeBURGER,
AND CHARLOTTE DENNY¹

University of Kentucky

The latest revision of the Stanford-Binet (S-B) test of intelligence was published in January 1960; therefore, it was thought advisable to compare the IQs of a group of white American children on the 1960 test with IQs made by the same group on the 1937 S-B and on the Wechsler Intelligence Scale for Children (WISC). In addition, Raven Progressive Matrices (1938 & 1947) and Goodenough Draw-A-Man (D-A-M) IQs for part of the group are compared with the two S-Bs and the WISC.

The group consists of pupils attending the University of Kentucky School, grades one through eight. Parents' socioeconomic status is above average; fathers' occupations are managerial and professional. Tested intelligence of the pupils is likewise above average; mean IQ on the 1960 S-B equals 123; the range is 84 to 159. The selection criterion was the availability of scores for the 1937 S-B L and M forms and WISC full scale. All pupils meeting this criterion were included in this study, making a total of 82 for the major comparisons, 47 boys and 35 girls.

The 1960 S-B was administered by the authors. All other tests were administered over a 4-year period by graduate students in psychology enrolled in an intelligence testing course. All of the tests were checked, rescored, and supervised by the senior author.

RESULTS

There were from one to four scores available for each test; therefore, means were used, when available, for greater reliability. The

1937 S-B score was a composite of the L and M scores.

Comparisons are reported in the form of group means and product-moment correlation coefficients.

The 1937 S-B scores were converted according to the equation provided in the 1960 S-B manual (Terman & Merrill, 1960, p. 339). This equation provides corrections for the mean and standard deviation which are not precisely equal for all chronological age groups for the 1937 S-B. The converted IQ scores were generally lower than the unconverted scores, mean deviations by intelligence level varying from two to four points. Overall, however, differences for means, standard deviations, and correlations were small and generally inconsistent and, therefore, comparisons involving the 1937 S-B are based on unconverted scores.

WISC, 1937 S-B, 1960 S-B

Age. Littell (1960), in a review of WISC studies, concluded that children who are younger and rank higher on the 1937 S-B tend to score higher on the 1937 S-B than on the WISC. Table 1 shows no differential discrepancies in the present study due to chronological age. The null effect of age has also been found by Weider, Noller, and Schramm (1951), Gehman and Matyas (1956), and Schacter and Apgar (1958); hence, it appears that the evidence for the age factor is inconclusive. When the 1960 S-B is compared with the 1937 S-B and with the WISC, again age is not found to be a factor.

¹ Denny is at the College of Nursing.

TABLE 1
IQ SCORES

Group	N	1960 S-B	1937 minus 1960	1937 S-B	1937 minus WISC	WISC	1960 minus WISC
Average	12	100	2	102	1	101	-1
High average	19	115	-1	114	3	111	4
Superior	34	126	2	128	9 ^a ± 1.24	119	7 ^b ± 1.92
Very superior	17	138	8 ^b ± 1.93	146	16 ^a ± 1.89	130	8 ^b ± 2.08
CA 6-10 to 10-0	33 ^c	123	1	124	8	116	7
CA 10-0 to 14-1	34 ^c	122	4	126	9	117	5
Entire group	82	123	2	125	7	118	5

Note.—Intelligence levels based on 1937 S-B scores.

^a Discrepancy = 0, $p < .00002$.

^b Discrepancy = 0, $p < .002$.

^c The N for the CA groups is less than for the entire group because some children changed from the lower to the higher CA classification during the 4-year testing period.

Intelligence. Littell's conclusion regarding the effect of IQ level on WISC-1937 S-B test discrepancies is supported in the present study. Two-tailed t tests were made when the discrepancies were greater than the five points usually allowed as test-retest error. For the two superior groups, the superiority of the 1937 S-B over the WISC is significant, $p < .00002$. This finding applies likewise to the 1960 S-B superiority over the WISC, $p < .002$. For average groups, WISC scores are comparable to both the 1937 and 1960 S-B scores. Discrepancies between both S-B tests and the WISC are greater for the superior levels than for the average levels, $p < .00002$ for the 1937 S-B and the WISC and $p = .02$ for the 1960 S-B and the WISC. Error terms for the 1937 S-B and the 1960 S-B comparisons are 1.61 and 2.13, respectively. 1937 and 1960 S-B scores are comparable except at the very superior level where the discrepancy is again highly significant, $p < .002$. Cronbach (1960, p. 171) reports, "IQs on the two scales are not strictly comparable."

Entire group. The major correlations for the entire group are presented in Table 2. Inter-correlations between the 1960 S-B, the 1937 S-B, and the WISC do not differ significantly from each other. Although the population is relatively small, the agreement found between the 1937 S-B and the WISC compares favorably with the correlations reported for white American children by Littell (1960), i.e., in the .80s. The correlation between the 1937 S-B and the WISC tends to be higher than that

between the 1960 S-B and the WISC. This might be expected for two reasons: (a) the test interval is greater for the 1960 S-B and WISC administrations and (b) the 1937 S-B scores are based on two to four tests and, hence, may be more reliable than the 1960 S-B scores which are based on one test. The correlation between the 1960 and 1937 S-B scores reaches a respectable validity figure. The 1937 L and M correlation compares favorably with that reported by Terman and Merrill (1937, p. 47) for the 1937 S-B standardization group.

Raven and Draw-A-Man

As a trend, the Raven Progressive Matrices is superior to the D-A-M in predicting the 1960 S-B, the 1937 S-B, and the WISC (Table 3). Differences between the Raven and D-A-M are not significant, however. While the Raven and D-A-M scores account for a significant amount of the variance of the three major tests, the magnitude of these correlations is relatively small, and, consequently,

TABLE 2
MAJOR CORRELATIONS

Test	1960 S-B	WISC	1937 S-B, Form L
1937 S-B	.82	.80—.80's	
WISC	.74		
1937 S-B, M			.92—.91

Note.—N = 82; all > 0, $p < .005$; representative finding on right.

TABLE 3
RAVEN AND DRAW-A-MAN CORRELATIONS

Test	1960 S-B	1937 S-B	WISC
D-A-M	.43	.46—.41	.43
Raven	.59	.67—.54	.55—.91 .75

Note.— $N = 72$; all > 0 , $p < .005$; representative findings on right.

individual predictability is low. The D-A-M and 1937 S-B correlation compares favorably with that found by McHugh (1945) for public school kindergarten children but is much lower than Goodenough's (1926, p. 51) D-A-M and 1916 S-B correlation of .74. The Raven and 1937 S-B correlation is higher than that found by Banks and Sinha (1951) for London school children, whereas the Raven and WISC correlation is much lower than the corresponding correlations of .91 and .75 found by Martin and Wiechers (1954) and Barratt (1956), respectively, for American school children. It is difficult to evaluate comparisons involving the Raven and D-A-M tests since there are so few representative studies in the literature.

DISCUSSION

The primary purpose of this report is to compare the 1960 S-B with its predecessor, the 1937 S-B, and another major children's intelligence test, the WISC. The 1960 S-B was found to correlate equally well with the 1937 S-B and the WISC, these correlations being comparable to representative findings for similar relationships. While the 1960 S-B was found to predict the 1937 S-B quite well for a fairly heterogeneous group, prediction was not equal for subgroups classified according to 1937 S-B intelligence levels. Agreement was found for average and superior levels but not for the very superior level. This partial disagreement of the two scales supports Cronbach's (1960) doubt that they are strictly comparable but further investigation with regard to intelligence level is indicated due to the relatively small size of the samples reported here.

Intelligence level was likewise found to be a significant factor preventing high agreement

between both S-B tests and the WISC, a fact well supported by comparisons of the 1937 S-B and the WISC. Implications from findings of the present study are that at average levels of intelligence WISC scores may be used interchangeably with scores from both S-B tests. This is not true for the superior levels where the obtained significant discrepancies of 7 to 16 points may easily place the scores in different intelligence classifications resulting in somewhat spurious or doubtful interpretations. However, it would be unwise to set definite limits for test comparability according to intelligence level on the basis of existing information which is not in complete agreement or strictly comparable. More information is needed drawn from larger and more representative samples. This information is no doubt available if existing test data were classified and analyzed.

SUMMARY

1. The comparability of IQs from five different intelligence tests was investigated for an above average group of white American children.

2. For the entire group ($N = 82$), scores for the 1960 S-B, the 1937 S-B, and the WISC were found to be comparable and to compare favorably with representative similar findings.

3. The age factor, contrary to some previous findings, was not found to account for test discrepancies among the two S-B and WISC instruments.

4. Intelligence level, in agreement with previous findings, was a factor producing highly significant discrepancies at superior levels among the two S-B and WISC instruments. More investigation is needed regarding the effect of intelligence level on test comparability.

5. Correlations relating the Raven and D-A-M to the two S-Bs and WISC were significant but relatively small.

REFERENCES

- BARRATT, E. S. The relationship of the Progressive Matrices (1938) and the Columbia Mental Maturity Scale to the WISC. *J. consult. Psychol.*, 1956, 20, 294-297.
- BANKS, CHARLOTTE, & SINHA, UMA. An item analysis of the Progressive Matrices Test. *Brit. J. Psychol.*, 1951, 4, 91-95.

- CRONBACH, L. J. *Essentials of psychological testing*. New York: Harper, 1960.
- GEHMAN, ILA H., & MATYAS, R. P. Stability of the WISC and Binet tests. *J. consult. Psychol.*, 1956, 20, 150-152.
- GOODENOUGH, FLORENCE L. *Measurement of intelligence by drawings*. New York: World Book, 1926.
- LITTELL, W. M. The Wechsler Intelligence Scale for Children: Review of a decade of research. *Psychol. Bull.*, 1960, 57, 132-156.
- McHUGH, GELOLO. Relationship between the Good-enough drawing a man test and the 1937 revision of the Stanford-Binet test. *J. educ. Psychol.*, 1945, 36, 119-124.
- MARTIN, A. W., & WIECHERS, J. E. Raven's Colored Progressive Matrices and the Wechsler Intelligence Scale for Children. *J. consult. Psychol.*, 1954, 18, 143-145.
- SCHACTER, FRANCES F., & APCAR, VIRGINIA. Comparison of preschool Stanford-Binet and school-age WISC IQs. *J. educ. Psychol.*, 1958, 49, 320-323.
- TERMAN, L. M., & MERRILL, MAUD A. *Measuring intelligence*. Boston: Houghton Mifflin, 1937.
- TERMAN, L. M., & MERRILL, MAUD A. *Stanford-Binet intelligence scale*. Boston: Houghton Mifflin, 1960.
- WEIDER, A., NOLLER, P. A., & SCHRAMM, T. A. The Wechsler Intelligence Scale for Children and the revised Stanford-Binet. *J. consult. Psychol.*, 1951, 15, 330-333.

(Received July 27, 1960)

EGO IDENTITY, ROLE VARIABILITY, AND ADJUSTMENT¹

JACK BLOCK

University of California, Berkeley

The meaning of the concept of "ego identity" is still evolving. In Erickson's perceptive essays on adolescent adjustment and behavior (1950, 1956), the construct is discussed and explored so richly that an easy and clearly sufficient operational translation of the notion cannot be formulated. A core meaning of the concept, however, is perhaps to be found in Erickson's (1950) statement that "the sense of ego identity is the (individual's) accrued confidence that (his) inner sameness and continuity are matched by the sameness and continuity of (his) meaning for others . . ." (p. 228).

In this definition, three elements are present. First, an individual must perceive himself as having "inner sameness and continuity," i.e., he must, over time, presume himself to be essentially the same person he has been. Second, the surrounding persons in one's social milieu must perceive a "sameness and continuity" in the individual also. And finally, the individual must have "accrued confidence" in a correspondence between the two lines of continuity, internal and external. His perception of the person he sees himself as being must be validated by feedback from his interpersonal experiences.

From this definition, one aspect of ego identity may be singled out for study, the dimension we have labeled *role variability* (RV). The meaning of role variability is perhaps most readily indicated by describing its extremes. At one end of this dimension, there is "role diffusion," where an individual is an interpersonal chameleon, with no inner core of identity, fitfully reacting in all ways to all people. This kind of person is highly variable in his behaviors and is plagued by self-doubts

and despairs for he has no internal reference which can affirm his continuity and self-integrity. At the other extreme, there is what might be called "role rigidity," where an individual behaves uniformly in all situations, disregarding the different responsibilities different circumstances may impose. Here the core of identity is hollow, based not on a genuine and unquestioned sense of personal integrity but rather upon deep seated fear of any amount of self-abandon. Somewhere in between, presumably, a proper balance can be struck in the struggle both for identity and the capacity for intimacy.

In terms of the preceding definition of ego identity, role variability may exist at the level of self-evaluation and separately, at the level of observations by others. It is the *relation* between these two levels of role variability that specifies ego identity or a problem of ego identity. In the present study, the focus has been limited to "sameness and continuity" as perceived and evaluated by the participant, and its significance for adjustment.

Some years ago, the writer reported a study of the way in which an individual's role behaviors changed as a function of various interactional contexts (Block, 1952). A single subject was studied by having her systematically describe her interactions with a set of "relevant others." These descriptions were then factor analyzed—an instance of O-technique (Cattell, 1946)—and it was observed that the factor dimensions appeared to order and to summarize, in a cogent way, the several kinds of roles this subject manifested.

In this earlier study, the focal subject viewed herself quite differently, depending upon the person with whom she was confronted, and yet she still presented a substantial core of interpersonal consistency. Although seeing herself as changing from rela-

¹ This investigation was supported by Research Grant M-1078 from the National Institute of Mental Health of the United States Public Health Service.

tionship to relationship, a general factor of some consequence proved to underlie all her interactions.

During the course of evaluation of this and other findings, the general question was raised, "How much . . . interpersonal consistency—i.e., interpersonal sameness—is socially appropriate and, in terms of the individual's internal psychic economy, consonant with his need systems?" (Block, 1952, p. 285). In Erickson's terms, role diffusion is a personally untenable situation for the individual. He is beset by too many despairs and self-contradictions as a consequence of his extreme behavioral fluctuations. Role rigidity, where an individual is not affected in his behaviors as a function of the persons with whom he is interacting, may be an indefinitely prolonged adaptation of sorts, but certainly is not an optimal interpersonal solution. It prohibits further growth and development of the individual; it is effective as a security mechanism only so long as the interpersonal surround is a tolerant one. From this frame of reference follows the simple hypothesis the present study endeavored to test: "the amount of interpersonal consistency is curvilinearly related to the degree of maladjustment, as defined independently" (Block, 1952, p. 285).

METHOD

Index of Role Variability

A set of 20 adjectives selected a priori as reflecting various fundamental facets of interpersonal behavior served as the basic descriptive device. Each subject ranked this set of 20 adjectives eight times, so as to characterize his own behavior while with each of eight specified "relevant others." The 20 adjectives employed were listed in the following order: relaxed, formal, indifferent, warm, independent, witty, cooperative, assertive, indecisive, distractible, humorous, insincere, masculine (or feminine if rater was a woman), wise, unselfish, trusting, worrying, gestic, responsive, and protective. The subjects were asked to describe via this ranking of adjectives technique their behavior and relationships with the following eight individuals: someone in whom you are sexually interested, an acquaintance you don't care much about, your employer or someone with equivalent status, a child, a parent (or parent figure) of the same sex, a parent (or parent figure) of the other sex, a close friend of the same sex, an acquaintance whom you would like to know better. These eight specified other persons were selected as sampling widely, if not exhaustively, the interper-

sonal possibilities of the subjects employed. The test materials were arranged in booklet form, the descriptions being recorded on separate pages.

For each subject, the eight adjective rankings were intercorrelated by the Spearman rank-difference correlation method and the resulting 8×8 correlation matrix factor analyzed by the Thurstone centroid method. For each matrix, the percentage of the total communal variance explained by the first unrotated factor was calculated. The first unrotated factor of a matrix reflects the degree of congruence among the set of variables being studied. For the present data, the first unrotated factor indicates the extent of interpersonal consistency a subject views himself as manifesting over the set of relevant others specified. By dividing the mean first factor loading (squared) by the average communality of the matrix, an index of role variability is derived which is comparable from matrix to matrix, and hence from individual to individual. When *high*, this score suggests that the individual involved views himself as essentially the same person in his several interactions—he is interpersonally consistent; when *low*, the subject has described himself as rather different from situation to situation—he is interpersonally changeable.² This interpersonal consistency score has a potential range of 0 to 100.

Measurement of Maladjustment

Each subject also responded to the 480-item California Psychological Inventory (CPI) (Gough, 1957). The CPI, by virtue of the conceptual scheme which has guided it from its inception, employs a pool of items which are understandable by and inoffensive to nonpsychiatric populations and it was for this reason this inventory was used, even though the established scales of the CPI provide no direct measure of maladjustment. Although a knowledgeable interpreter can readily discern maladjustment from the nature of a subject's CPI profile, for our present purposes a more direct CPI measure of maladjustment was desired.

As part of the continuing program of scale development at the Institute of Personality Assessment and Research, it proved possible several years ago to establish, refine, and validate a personality scale to index an individual's "susceptibility to anxiety." This scale, labeled Psychoneuroticism (P_n), was developed by the sequential application of cluster analysis, item analysis against cluster score criteria, a further method of dimensional purification, and finally, validation on a number of different subject samples. Another paper will describe in some detail the development of the P_n scale and several other new scales

² Often, extracting the first factor from each subject's interperson matrix will prove to be too endless a job of calculation for the individual researcher. In such instances, Kendall's coefficient of concordance, computed for each subject, may be substituted as a quick and largely adequate index of interpersonal consistency.

with unusual properties but for the present report a brief description of the *Pn* scale is in order.

The *Pn* scale contains a total of 45 items³ of which 33 are contained in the MMPI item pool and 19 in the CPI collection of items (the apparent inconsistency in arithmetic here is due to the inclusion in the CPI of a number of MMPI items). The 19 *Pn* items scorable from a CPI protocol are sufficient in number to provide a reliable and dimensionally valid score.

The psychological meaning of the *Pn* scale is best conveyed for the purposes of this paper by a listing of its relationships to other established scales. It correlates in the .70s and sometimes .80s with the MMPI Psychasthenia scale, the MMPI Manifest Anxiety scale compiled by Taylor (1953), and the Anxiety scale developed from a factor analysis by Welsh (1956). All of these scales are good representatives of the first underlying dimension repeatedly found in factor and cluster analyses of personality inventories (cf., e.g., Block & Bailey, 1955b; Cook & Wherry, 1950; Cottle, 1950; Kassebaum, Couch, & Slater, 1959; Tyler, 1951; Wheeler, Little, & Lehner, 1951). This dimension, variously measured, has proved of broad significance in both correlational and experimental studies (cf., e.g., Block & Bailey, 1955a; Eriksen, 1954; Farber & Spence, 1953; Taylor, 1956) and it is clear now that individuals at the unfavorable end of the continuum tend to be troubled, self-preoccupied, and vulnerable to happenings that for most would go unnoticed. The choice, in particular, of the *Pn* scale to index this dimension of susceptibility to anxiety was in large part dictated by the option it provided of scoring individuals from CPI protocols. This choice was buttressed in addition by the scale's factorial origins and its development on nonpathological samples.

Subjects

Forty-one college students in a class on factor analysis collected the interpersonal and CPI data (and factored their respective small matrices). Anonymity was preserved for the participants. Many of the students collecting the data later admitted having used themselves as subjects in order to test the psychological insightfulness of the later factor analytic results. For this reason, it seems likely that the responses of the subjects to both the interpersonal consistency and CPI procedures were offered with appropriate motivation.

³ A complete list of the CPI and MMPI items defining the *Pn* scale has been deposited with the American Documentation Institute. Order Document No. 6828 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance \$1.25 for microfilm or \$1.25 for photocopies. Make checks payable to Chief, Photoduplication Service, Library of Congress. Copies of the scale are also available from the writer upon request.

RESULTS

For the measure of interpersonal consistency, the mean score was 71.42, the standard deviation being 13.73. The distribution of scores was somewhat negatively skewed. The mean score on the *Pn* scale was 7.35, with a standard deviation of 3.02. Here the score distribution was moderately skewed positively.

The hypothesis originally advanced asserts that a significant curvilinear relationship, as measured by eta, should exist between interpersonal consistency and maladjustment. A scatter plot, however, shows quite clearly that a linearity assumption fits the data well. The product-moment correlation between the index of interpersonal consistency and scores on the *Pn* scale is $-.52$, significant beyond the .001 level. Individuals who tend to see themselves as varying from interaction to interaction are also more maladjusted, as measured by the *Pn* scale. The expectation that individuals with too little role variability would also prove to have weaknesses in their personality makeup was not confirmed.

In order to understand more closely the significance of the interpersonal consistency index, an analysis was undertaken of the 480 items in the CPI. The 20 individuals with the highest interpersonal consistency score were constituted as one group and the 20 individuals with the lowest interpersonal consistency scores were formed into a second group. For each of the CPI items, the relative frequencies of response of the two groups were evaluated by means of Fisher's exact test for 2×2 contingency tables. The number of items discriminating beyond the .05 level of significance was in excess of the number to be expected on the basis of chance by a factor of three, clear evidence for their nonchance nature (Block, 1960). The content of these discriminating items can thus serve to enrich our understanding of the psychological meaning of the index of interpersonal consistency. To bring some order into the set of distinguishing items, they are presented as grouped into four tentative and somewhat overlapping categories. Except where noted, for all the items listed, the individuals who are changeable in their interpersonal role tend to say "Yes" more often. Significance level and frequencies

of the Yes response in the interpersonally changeable and interpersonally consistent groups are indicated in parentheses.

Items expressing social inarticulateness and concern

83. I usually feel nervous and ill at ease at a formal dance or party (.01, 9-1).

134. It makes me uncomfortable to put on a stunt at a party even when others are doing the same sort of thing (.01, 9-0).

173. My way of doing things is apt to be misunderstood by others (.05, 8-1).

198. Before I do something I try to consider how my friends will react to it (.05, 14-6).

200. (Affirmed more by interpersonally consistent individuals) In a group of people I would not be embarrassed to be called upon to start a discussion or give an opinion about something I know well (.05, 11-18).

260. (Affirmed more by interpersonally consistent individuals) I always try to do at least a little better than what is expected of me (.05, 9-16).

285. I refuse to play some games because I am not good at them (.01, 12-3).

373. My table manners are not quite as good at home as when I am out in company (.01, 20-13).

Items expressing personal tension and neurotic character

358. I dream frequently about things that are best kept to myself (.05, 5-0).

406. I have one or more bad habits which are so strong that it is no use fighting against them (.05, 5-0).

425. I have often felt guilty because I have pretended to feel more sorry about something than I really was (.05, 7-1).

453. I work under a great deal of tension (.05, 7-1).

Items expressing cynicism based upon disappointment

48. Most people would tell a lie if they could gain by it (.01, 15-5).

219. Most people inwardly dislike putting themselves out to help other people (.05, 8-1).

225. People pretend to care more about one another than they really do (.05, 12-4).

342. Some people exaggerate their troubles in order to get sympathy (.05, 19-13).

446. I must admit that people sometimes disappoint me (.05, 20-14).

Items expressing familial tension

164. My parents have often disapproved of my friends (.01, 11-0).

268. At times, I have been very anxious to get away from my family (.05, 19-12).

271. My parents were always very strict and stern with me (.01, 8-0).

An item not admitting categorization

333. Education is more important than most people think (.05, 19-13).

Summarizing the sense of these discriminating statements, it seems clear that the individuals seeing themselves as highly variable in their interactions experience strain and dismay in their social endeavours, view the world as essentially unfriendly, and are personally troubled.

DISCUSSION

The results obtained support but one slope of the hypothesized inverted U relationship. Although extreme role variability does appear to be related to an independent index of personality maladjustment, extreme role stability, at least as represented in this study, is not also indicative of neurotic qualities. This partial support of the hypothesized relationship, moreover, is not a one time finding for another, albeit less adequate, test of the same hypothesis derived equivalent results. In a study of 50 Vassar alumnae some 20 years after graduation, where interpersonal inconsistency was defined in almost exactly the way already indicated, role stability correlates again showed a consistent picture of (relative) psychological health. For example, women who are stable in their interpersonal role were described in ratings formulated completely independently of scores on interpersonal consistency as relatively "indulgent and forgiving, protective of those close to her, sympathetic, efficient, adequate in her sexual role, turned to for advice and reassurance, facially and gesturally expressive, and considerate." Women who are variable in the interpersonal behavior were described as "irritable and overreactive, talkative, ostentatious, and sarcastic." Interpersonal consistency correlated .29 with a consensus rating of degree of adjustment, a relationship significant at the .05 level. A finding by Meltzer (1957) to the effect that a large self-ideal-self discrepancy—a reasonable measure of self-recognized maladjustment—is significantly related to extreme role variability, defined by means equivalent to those used here, may also be interpreted as congruent with the present findings. It seems fair to conclude, then, that a positive relation exists between role variability as indexed here and the kind of maladjustment where the individual explicitly and consciously acknowledges his personal vulner-

ability. There is as yet no empirical suggestion that extreme interpersonal consistency is also related to maladjustment.

When a hypothesis achieves partial but not complete confirmation, there is the possibility that either the hypothesis is wrong as stated or that it was not tested properly. Before abandoning a hypothesis held likely on other grounds, it is required that the measures employed be evaluated for their adequacy and the sample in which the relationship was sought be evaluated with reference to its appropriateness for the hypothesis being tested. Let us consider these in turn.

The use of a personality scale such as *Pn* to index the dimension variously called "susceptibility to anxiety," "neuroticism," "ego weakness" and so on is, as referenced earlier, both well-established and well-supported. In a large number of studies, the correlates of scores reflecting this dimension have testified to its importance and meaning. We may presume, therefore, that the maladjustment index employed in the present study is both representative and, on the whole, effective.⁴

The merits and properties of the interpersonal consistency measure of course cannot be fully evaluated presently. In support of its construct validity, however, a number of arguments can be adduced. The operations appear to parallel the conceptual steps involved in deriving the notion of role variability. The task presented to subjects is, on the face of it, not especially threatening, hence removing much of the conscious motivation for offering "safe" and uninformative responses. The way in which the subject's data are then processed so as to provide a score is sufficiently complicated and removed as to prevent a subject from readily controlling, when he ranks his

adjectives, the score he later obtains. Finally, the theoretically appropriate if as yet insufficient correlations of the measure with independent variables in several studies suggest its proximity, at least, to the underlying concept of role variability.

Perhaps the primary reason why the originally formed hypothesis failed of complete confirmation in this and the other studies cited is that the samples involved, in all three cases, may have been too small and too homogeneous to contain enough individuals who, in an absolute sense, were "role rigid." This is a post hoc explanation, of course, but it may be that individuals who go on to college, in the natural course of their selective evolution, necessarily develop *some* amount of flexibility in their role behaviors. College students cannot be insensitive and unresourceful before the various role demands made upon them and therefore, in working with college samples, very many of the individuals who are rigidly the same in their interpersonal endeavors may already have been screened out.

As another alternative to explore, it may be that extreme interpersonal consistency is an aspect of personality at a somewhat later age, when the personally desperate individual has found a self-definition that is acceptable to him. This last possibility is infirmed somewhat by the inability to identify such individuals in the older Vassar sample. What is required is another study, this time of a much larger and preferably less homogeneous sample so that individuals who are extremely stable in their role behaviors may fairly be presumed to have been included. Although there has been confirmation and cross-validation of the hypothesis that extreme role variability will be associated with personality maladjustment, it is premature, we would suggest, to abandon the additional hypothesis that extreme interpersonal consistency is also associated with personality maladjustment. A further empirical effort is needed to discover whether there is a far side to the mountain. In the meanwhile, to the extent that role variability as measured here relates to role diffusion as conceived by Erickson, the present study offers support for the implications he has drawn between ego identity and behavior.

⁴ It is important to note that the *Pn* scale is, in this and a number of other studies, orthogonal to the Ego Control (*EC*) Scale, a scale developed to measure tendency to constrict or to express impulse. The *EC* scale correlates an insignificant .15 with role consistency, reflecting a slight tendency for over control to go along with reduced role variation. Although certain scales reasonably equivalent to *Pn* emphasize expressive or overt reactions to anxiety and de-emphasize suppressive and indirect reactions, the interrelations of *Pn*, *EC*, and the role consistency measure in the present study suggest that the failure of *Pn* to be related to role rigidity cannot be ascribed to a deficient representation of covert maladjustment.

SUMMARY

From Erickson's concept of ego identity, the dimension of role variability was abstracted. The hypothesis was advanced that excessive role variability ("diffusion") and insufficient role variability ("rigidity"), because they both reflect problems in ego identity, would both be associated with maladjustment. A measure of the extent to which an individual perceives himself as varying in a variety of interpersonal situations was developed. Role variability, so measured, proved to relate significantly to maladjustment, as measured by a CPI scale to measure "susceptibility to anxiety." Role rigidity did not relate to maladjustment. Supplementary findings were introduced and some possible reasons for the only partial confirmation of the curvilinear hypothesis were offered.

REFERENCES

- BLOCK, J. The assessment of communication: Role variations as a function of interactional context. *J. Pers.*, 1952, 21, 272-286.
- BLOCK, J. On the number of significant findings to be expected by chance. *Psychometrika*, 1960, 25, 369-380.
- BLOCK, J., & BAILEY, D. E. *A cluster analysis of 82 inventory measures of personality, interest, and intellect*. Berkeley, Calif.: Institute of Personality Assessment and Research, 1955. (a)
- BLOCK, J., & BAILEY, D. E. Q-sort item analyses of a number of MMPI scales. Technical Memorandum OERL TM-55-7, May 1955, Maxwell Air Force Base, Alabama. (b)
- CATTELL, R. B. *The description and measurement of personality*. New York: World Book, 1946.
- COOK, E. B., & WHERRY, R. J. A factor analysis of MMPI and aptitude test data. *J. appl. Psychol.*, 1950, 34, 260-265.
- COTTLE, W. C. A factorial study of the multiphasic, Strong, Kuder, and Bell inventories using a population of adult males. *Psychometrika*, 1950, 15, 25-47.
- ERICKSON, E. H. *Childhood and society*. New York: Norton, 1950.
- ERICKSON, E. H. The problem of ego identity. *J. Amer. Psychoanal. Ass.*, 1956, 4, 56-121.
- ERIKSEN, C. W. Psychological defenses and "ego strength" in the recall of completed and incomplete tasks. *J. abnorm. soc. Psychol.*, 1954, 49, 45-50.
- FARBER, I. E., & SPENCE, K. W. Complex learning and conditioning as a function of anxiety. *J. exp. Psychol.*, 1953, 45, 120-125.
- GOUGH, H. G. *The California Psychological Inventory*. Palo Alto, Calif.: Consulting Psychology, 1957.
- KASSEBAUM, G. G., COUCH, A. S., & SLATER, P. E. The factorial dimensions of the MMPI. *J. consult. Psychol.*, 1959, 23, 226-235.
- MELTZER, M. L. Role variability as a function of the understanding of others. Unpublished doctoral dissertation, Catholic University of America, 1957.
- TAYLOR, JANET A. A personality scale of manifest anxiety. *J. abnorm. soc. Psychol.*, 1953, 48, 285-290.
- TAYLOR, JANET A. Drive theory and manifest anxiety. *Psychol. Bull.*, 1956, 53, 303-320.
- TYLER, F. T. A factorial analysis of fifteen MMPI scales. *J. consult. Psychol.*, 1951, 15, 451-456.
- WELSH, G. S. Factor dimensions A and R. In G. S. Welsh & W. G. Dahlstrom (Eds.), *Basic readings on the MMPI in psychology and medicine*. Minnesota: Univer. Minnesota Press, 1956.
- WHEELER, W. M., LITTLE, K. B., & LEHNER, G. F. The internal structure of the MMPI. *J. consult. Psychol.*, 1951, 15, 134-141.

(Received August 1, 1960)

SOMATIC EXPERIENCE IN THE ANXIETY STATE: SOME SEX AND PERSONALITY CORRELATES OF "AUTONOMIC FEEDBACK"¹

SHELDON J. KORCHIN² AND HELEN A. HEATH

Michael Reese Hospital, Chicago, Illinois

One facet of the anxiety state is the experienced alteration of somatic functioning. Recent work by Mandler and his associates (Mandler & Kremen, 1958; Mandler, Mandler, & Uviller, 1958) has again called attention to this aspect of anxiety, termed by them "autonomic feedback." An Autonomic Perception Questionnaire (APQ) was developed for the self-description of somatic symptoms characteristic of subjects' anxiety experience. They have investigated whether and how the number and/or intensity of such symptoms is related to other measures of anxiety and to autonomic reactivity under stress as measured directly. In their first study, which compared selected high and low APQ scorers, high subjects more commonly reported somatic sensations during intellectually stressful tasks, and generally showed greater autonomic reactivity in terms of polygraphic measurements (Mandler et al., 1958). While generally replicating these findings, a second study of unselected subjects found less distinct relationships between APQ scores and either somatic report or autonomic reactivity under the same stress (Mandler & Kremen, 1958). Correlations between the APQ and both the Taylor Manifest Anxiety scale and a newly developed Body Perception Scale were positive in both studies,

though considerably lower in the second. The present study is designed to replicate and extend these findings, and generally to explore further the personality correlates of autonomic feedback.³

METHOD

The Nowlis Adjective Check List (ACL), Barron Ego Strength scale (*Es* scale), Taylor Manifest Anxiety scale (*MA* scale), and the Autonomic Perception

³ For consistency with the earlier work, Mandler's terms will be used in this paper. However, the more inclusive and neutral term "somatic experience" seems to describe better the range of phenomena included in Mandler's "autonomic perception" or "autonomic feedback." Apparently, Mandler intended by these terms to describe sensitivity to the various bodily sensations which subjects might describe in emotional states, regardless of whether they reflect autonomic nervous system activity as such. Certainly, different orders of neurological and physiological control are suggested by such diverse APQ items as:

7. When you feel anxious, are you aware of increased muscle tension?

8. . . . do you get a headache?

16. . . . do you feel as if blood rushes to your head?

19. . . . do you get a sinking or heavy feeling in your stomach?

Just as "autonomic" suggests a too specific physiological mechanism, the words "perception" and "feedback" connote too immediate and definite a relationship between the physiological event and the experienced symptom. Precisely how experienced symptoms are related to measurable physiological activity is the research question at the core of Mandler's concern, though not an issue in this report. But, for clarity, it should be remembered that APQ assesses the variety and intensity of self-reported bodily experience, and that it defines a variable akin to what in clinical contexts is described as somatization or even hypochondriasis. Indeed, it would be appropriate and perhaps clearer to distinguish high and low APQ scorers as "somatizers" and "nonsomatizers."

¹ This study was supported in part by the Mental Health Fund of the State of Illinois and in part by Grant M-1442 of the National Institute of Mental Health. The authors are grateful to John Dubocq, Dean of Students at George Williams College, for his considerable interest and help in arranging time and facilities and in engaging the cooperation of his students and colleagues. Our thanks, too, to Barbara White for help in test administration and scoring.

² Now at the National Institute of Mental Health, Bethesda, Maryland.

Questionnaire (APQ) were administered, in that order, to better than half of the entire student body of George Williams College in Chicago. Each of the four college classes was tested in a separate session. The study, it was explained, was intended to explore some relations among test measures of emotional states. Students were assured that individual scores would be confidential, and that only group results might be discussed with their college authorities. The present report is based on the scores of 176 subjects (139 men and 37 women) for whom complete protocols were available. Subjects ranged in age from 16 to 40 years.

Mood Adjective Check List. The list consists of 140 adjectives (including 10 duplicates) which the subject rates on a four-point scale as more or less characteristic of his mood state (Nowlis, 1953). Factor analytic studies reported by Nowlis and Green (1957; Green & Nowlis, 1957) reveal eight factors: A, Concern; B, Aggression; C, Pleasantness; D, Activation-Deactivation, a bipolar factor; E, Ego-tism; F, Social Affection; G, Depression; and H, Anxiety. The original instructions request that the subject reply in terms of his *present* mood. In this study, he was asked instead to judge the items in terms of what is *generally characteristic* of him, since our concern is less with the subject's momentary mood than with his perception of more enduring modes of emotional behavior. As an additional rough estimate of emotional lability, which will not be treated in this report, the subject was also asked to circle all adjectives which described feelings he had during the preceding 24 hours.

Barron Ego Strength scale. This scale was derived empirically from the MMPI in terms of items which distinguished patients who benefited from psychotherapy from those who did not (Barron, 1953, 1954). From correlations with other assessment measures in additional patient and normal samples, Barron interprets the test as a measure of ego strength, including such characteristics as health and physiological stability, strong reality sense, feelings of adequacy, vitality, lack of prejudice, emotional spontaneity and outgoingness, and intelligence.

Taylor Manifest Anxiety scale. This procedure is also a derivative of the MMPI, consisting of items originally selected by clinical psychologists as exemplifying the manifest symptoms of anxiety (Taylor, 1953). The Taylor and Barron scales were administered in a combined form.

Autonomic Perception Questionnaire. The complete questionnaire described by Mandler, Mandler, and Uviller (1958) was administered and scored according to their procedure. The most relevant portion, for the present study, consisted of 21 graphic scale items which comprise their "Anxiety APQ" score. Each of these items starts with the dependent clause "When you feel anxious . . ." and then inquires into the frequency and/or intensity of symptoms in each of seven areas of bodily function. For example, the item "When you feel anxious, do your hands become cold?" is rated on a scale from "No change" to "Very cold." An additional nine items describe symptoms

TABLE 1

SEX DIFFERENCE IN MEAN AUTONOMIC PERCEPTION QUESTIONNAIRE, TAYLOR MANIFEST ANXIETY SCALE, AND BARRON EGO STRENGTH SCALE

Test	Men	Women	<i>t</i>
Anxiety APQ	66.30	81.00	2.726*
Pleasure APQ	16.96	19.46	1.211
MA scale	12.66	17.65	3.698**
Es scale	48.95	46.43	2.620*

* $p < .01$.

** $p < .001$.

related to pleasure. These are of the same form, though starting with the clause "When you feel happy . . ." The sum of these nine items provides a "Pleasure APQ," which will be of lesser concern in this report.

An additional group of nine anxiety items was developed to extend coverage in the areas sampled, and to tap further somatic areas (e.g., faintness, nausea, polyuria, diarrhea). Scores based on these new items correlated highly with the original, 21-item, Anxiety APQ— $r = .736$ (male subjects) and $r = .739$ (female subjects), both significant at $< .001$ level. Moreover, the new and original APQ scores vary in precisely the same way with each of the other personality measures. However, for greater comparability with earlier work, the Anxiety APQ analyses reported in this paper are based only on the original 21 items.

RESULTS AND DISCUSSION

Sex Differences in Mean Test Scores

There are distinct and significant differences in mean APQ, MA scale, and Es scale scores between men and women (Table 1). Women, in this population, are higher in manifest anxiety, lower in ego strength, and report more somatic experience. The female mean Pleasure APQ is also higher, though insignificantly so.

The magnitude of these differences is puzzling, particularly since none of the original reports of these procedures describes similar sex differences in seemingly comparable populations. Although women had slightly higher MA scale values in the original Iowa sample (Taylor, 1953), the difference was insignificant. Mandler and Kremen (1958) found no difference in mean APQ between Harvard summer school men and women, though they did discover some differences in autonomic response measures. Similarly, Barron (1953)

TABLE 2

INTERCORRELATIONS AMONG AUTONOMIC PERCEPTION QUESTIONNAIRE, TAYLOR MANIFEST ANXIETY SCALE, AND BARRON EGO STRENGTH SCALE FOR MALE AND FEMALE SUBJECTS

Test	Anxiety APQ		Pleasure APQ		MA scale		Es scale	
	M	F	M	F	M	F	M	F
Anxiety APQ			.529*	.509*	.410*	.634*	-.258*	-.423*
Pleasure APQ					.167	.288	-.300*	-.317
MA scale							-.590*	-.784*

Note.—Male $N = 139$; female $N = 37$.

* $p < .01$.

mentions no sex difference in ego strength scores.

In view of the decided sex differences in these measures, it is interesting that sexes do not differ at all in the ACL factor scores for anxiety and depression. Indeed, only one of the nine ACL scores (Social Affection, $p < .05$) is significantly higher for women than for men. There is some tendency for women to be higher in Activation, but only at the $p < .10$ level.

There is no ready explanation for the sizable sex differences. However, in view of them, it seems necessary to make further correlative analyses of tests within each sex group separately, lest the pattern of intertest correlation also differ between sexes. It should be noted, at this point, that the pattern of mean differences—higher APQ and MA scale going with lower Es scale—is in the direction predictable from the pattern of intertest correlation.

Relationship of Autonomic Feedback to Manifest Anxiety and Ego Strength

Intercorrelations among APQ, MA scale, and Es scale for male and female subjects separately are given in Table 2. For both sexes, it is clear that those who report more somatic experiences in emotional states are higher in manifest anxiety and lower in ego strength. The MA scale and Es scale, as might have been anticipated, show substantial negative correlations. The correlation pattern supports Mandler's contention that autonomic perception is part of the anxiety complex. Subjects who are less capable of integrative functioning and who are simultaneously more prone to emotional disturbance experience a wider range and more intense somatic symp-

toms. The APQ vs. MA scale correlations of this study are of about the same order as those reported by Mandler et al. (1958) in their first report, though somewhat higher than the coefficient of .267 between Anxiety APQ and MA scale reported in their second study (Mandler & Kremen, 1958). Although the correlations between procedures within each sex group are generally in the same range, it is noteworthy that those for women are in all cases higher.

Relationship of Autonomic Feedback to Adjective Self-Description

In order to compare autonomic feedback to subjects' ACL self-description, the male and female distributions of Anxiety APQ scores were divided into near-equal thirds; thus forming high, median, and low APQ groups of men and women. These APQ groups are compared, first, in terms of the Nowlis-Green factor analytically defined variables (Table 3) and, second, in the distribution of their ratings on the 130 adjectives considered individually (Table 4). Three-way split of the APQ distribution was used to detect curvilinearity, if present. Although the mean ACL scores of the median APQ group often fell close to, or beyond, one of the extreme groups, in none of the significant comparisons presented in Tables 3 and 4 was the relationship clearly curvilinear. Hence, these findings may be taken as generally descriptive of the differences in ACL variables between higher and lower APQ scorers, of a roughly linear sort and certainly characteristic of extreme APQ groups.

The two factor variables which best distinguish the APQ groups are Anxiety and De-

pression (Table 3). Among both men and women, high APQ subjects are significantly higher in their mean Anxiety and Depression scores than the median and low APQ scorers. Suggestive, though less impressive, differences are found in the sets of scores defining the two poles of the Activation factor. Both male and female high APQ subjects tend to describe themselves in Deactive terms (though the F ratio is not significant for women). However, among women, high APQ scores also tend to be higher in Activation, while the men if anything trend in the opposite direction.

More detailed examination of the individual adjective ratings amplifies these findings, and makes clearer the apparent sex difference in Activation. In Table 4 are listed the adjectives in which the three APQ groups differ, on chi square analysis, at five levels of significance. It might be noted that one would expect by chance 26 significant comparisons at the .10 level or better in the 260 comparisons (130 adjectives, 2 groups). In fact, there are 56, over twice as many. Of greater importance, however, is the internal consistency and apparent sense that can be made from the pattern of adjective self-descriptions.

High APQ *male* subjects, contrasted to those reporting fewer and less intense somatic symptoms, describe themselves as inadequate, inactive, and helpless people. The self-image of weakness and incompetence is conveyed by the positive endorsement of such items as

weak, helpless, hesitant, and the like; and by low ratings on items which suggest active mastery, such as independent, resourceful, and effective. Emotionally, the adjective pattern suggests depression and feelings of futility—ashamed, downhearted, clutched up, frustrated—rather than any acute emotional distress such as might be expressed in hostility or anxiety. They are defeated, rather than angry men. Overall, one has the impression of ego-weak, inadequate, and dependent people whose symptoms are more those of "neurotic debility" than of acute emotional distress.

While high APQ *female* subjects, compared to their low APQ sex mates, also convey a distinctly neurotic impression, the pattern of adjective self-description differs from the male and suggests a more complex syndrome. As with the men, there are signs of inadequacy and weakness, but along with this considerably more evidence of stronger and more labile emotions—belligerent, lonely, overjoyed, irritated, angry. These subjects seem to experience higher levels of excitement, with wider swings of mood and activity, and more capacity for energetic striving, though perhaps with no more assurance of success than the men. Compared to their male counterparts, high APQ females seem less concerned with their inadequacy, while describing more hostile and generally emotional interaction. They are more active and aggressive, men more passive and self-doubting. In terms of

TABLE 3

COMPARISON OF NOWLIS ADJECTIVE CHECK LIST FACTOR SCORES FOR LOW, MEDIAN, AND HIGH AUTONOMIC PERCEPTION QUESTIONNAIRE SCORERS, MALE AND FEMALE SUBJECTS SEPARATELY

	Male APQ groups					Female APQ groups				
	Low	Median	High	F	p	Low	Median	High	F	p
Concentration (A)	8.5	8.0	8.1	.94	ns	7.8	8.2	8.2	.13	ns
Aggression (B)	5.6	5.0	5.8	1.86	ns	3.8	4.4	6.1	2.08	ns
Pleasantness (C)	7.9	7.5	7.7	.68	ns	7.4	8.1	8.8	1.37	ns
Activation (D+)	9.5	9.1	8.6	1.62	ns	8.7	9.9	10.8	2.95	<.10
Deactivation (D-)	4.5	4.3	5.6	3.22	<.05	4.2	4.2	5.8	2.02	ns
Egotism (E)	4.1	3.9	4.7	1.70	ns	3.5	3.0	4.3	1.22	ns
Social Affection (F)	9.3	9.2	9.0	.64	ns	9.8	9.9	10.4	.32	ns
Depression (G)	3.5	3.8	5.1	5.52	<.01	2.2	2.7	6.1	8.82 ^a	<.01
Anxiety (H)	3.5	3.0	4.4	6.35	<.005	3.1	3.2	5.2	5.52	<.01

Note.—Male $N = 139$; female $N = 37$.

^a Variances were not homogeneous; therefore, the Kruskal-Wallis H test was substituted for the F test.

TABLE 4

ADJECTIVE SELF-RATINGS WHICH SIGNIFICANTLY DIFFERENTIATE HIGH, MEDIAN, AND LOW ANXIETY AUTONOMIC PERCEPTION QUESTIONNAIRE SCORERS WITHIN MALE AND WITHIN FEMALE GROUPS

Significance level	Adjectives* which differentiate high, median, and low APQ among:	
	Men	Women
<.001		<i>jittery</i> <i>doubtful</i> <i>shocked</i>
<.01	<i>helpless</i> <i>inactive</i> <i>thirsty</i> — wideawake	<i>insecure</i> full of pity sleepy belligerent lonely
<.02	ashamed washed out grouchy — independent — effective	<i>downhearted</i> regretful overjoyed — calm
<.05	<i>downhearted</i> <i>startled</i> <i>frustrated</i> <i>hesitant</i> slow bored clutched-up weak — resourceful — bold — careful	<i>helpless</i> <i>startled</i> energetic skeptical irritated angry — optimistic
<.10	<i>self-conscious</i> <i>jittery</i> <i>insecure</i> <i>dissatisfied</i> timid smug blue — serious — alert — satisfied	<i>self-conscious</i> <i>frustrated</i> <i>hesitant</i> careless boastful subdued restrained

Note.—Male $N = 139$; female $N = 37$.

* Adjectives are recorded here exactly as they appear in the ACL. If there is an inverse relation between APQ level and adjective rating (i.e., Highs rated lower than lows), a minus sign appears before the adjective. Adjectives which appear in both the male and female list are italicized.

the sex norms of our culture, one might speculate that the core neurotic problem expressed by each high APQ group represents inability to meet its particular sex standards: the males

are too dependent and ineffectual to be men; the females are too hostile and energetic to be women. The two adjective patterns suggest the "castrated male" and "penis-envy female" syndromes of psychoanalysis. Less speculatively, it can be concluded that ACL analyses fully support the earlier presented correlative analyses in showing individuals reporting greater autonomic feedback to be generally more neurotic, less well integrated, and more prone to emotional disturbance, while further suggesting personality characteristics at odds with effective sex role functioning.

Somatic Symptom Choice

It may be of some general interest to note the relative popularity of various somatic symptoms contained in the APQ, and to consider whether the sexes differ in their symptom choice. For rough and exploratory analysis, the 30 item means (original 21 plus our additional 9 items) were separately ranked for women and men, to compensate for the sex differences in mean scores already noted.

In general, men and women are quite similar in their relative orders of symptom choices. Below are listed the five items receiving the highest ranks and the five ranked lowest by the entire student group. In parentheses, the rank for men, then for women, is indicated. Those items added to the original APQ are marked with an asterisk. Recall that all items are preceded by the stem, "When you feel anxious. . . ."

Most highly rated items:

- do you get a fluttering feeling in your stomach ("butterflies")?* (1, 1)
- how often are you aware of bodily reactions? (3, 2)
- are you aware of increased muscle tension? (2, 6)
- do you ever feel weak or shaky?* (6, 3)
- are you aware of many bodily reactions? (4, 7)

Least highly rated items:

- do you experience nausea?* (30, 30)
- do you feel as if you might faint?* (29, 27)
- do you have to defecate frequently (diarrhea)?* (28, 28)

do you experience a slowing of the heart?*

(27, 29)

do you get a headache? (26, 23)

Note that four of the five least common symptoms are in response to items introduced by us. Perhaps such symptoms occur only with higher degrees of anxiety than these essentially healthy young men and women have experienced.

Seven of the items had identical ranks, or ranks differing only by one unit. In addition to the four included in the lists above, men and women were essentially alike on:

does your mouth become dry? (10, 10)

do you have to urinate frequently? (15, 16)

are you bothered by your bodily reactions?

(17.5, 18.5)

By contrast, inspection of the seven items which most differed in the rankings for men and women suggests some sex differences:

do your hands become cold? (25, 8.5)

do you have difficulty talking? (19, 8.5)

does the intensity of your heart beat increase? (11, 20)

how often are you aware of change in your breathing? (12, 21)

does your stomach get upset? (20, 13)

do you perspire? (5, 11.5)

do you get a lump in your throat or choked-up feeling? (9, 4)

Though the overall picture is one of greater agreement than difference in the somatic ways in which men and women express anxiety, the differences that are found seem consistent with clinical psychiatric and psychosomatic experience. Thus, women more commonly name cold hands, reminiscent of Raynaud's disease which is more frequently found among women, and difficulty in talking and lump in the throat, suggestive of hysteria. Though the differences are somewhat less great, men favor cardiac and respiratory symptoms, perspiration, and muscle tension, which taken together suggest the normal physiological reactions to exercise and exertion. Cardiovascular diseases are more commonly found among men, at least as indicated by mortality statistics in middle age. An important and long standing question in psychosomatic research, to which these data

may contribute slightly, is whether there are individual and group, such as sex, differences in somatic response in the acute emotional state which might be predictive of "symptom choice" in the psychosomatic disease state.

Relation between Anxiety and Pleasure APQ

Thus far, attention has been centered on somatic experience in anxiety (the anxiety APQ score), and consideration of the significant positive correlation between Anxiety and Pleasure APQ noted in Table 2 has been bypassed. In their original study, Mandler, Mandler, and Uviller (1958) report correlations of .50 and .45, for two male samples, between the two autonomic feedback scores. Although little systematic attention is given to this relationship—and in their later report Mandler and Kremen (1958) do not treat Pleasure APQ at all—these findings suggest that somatic sensitivity may be an individual characteristic in all or many states of emotional arousal. Generally anxious and ego-weak people, as we have already seen, are more prone to such experience, but apparently when happy as well as when anxious. At the same time, all subjects report more somatic experience in anxiety than in pleasure. Comparison of the duplicated items in the two scales shows consistently and significantly higher ratings for the Anxiety than for the Pleasure form. Apparently, then, while all subjects experience more somatic involvement in anxiety, individual differences in sensitivity to such experience are consistent across both states.

SUMMARY

This paper reports a correlative study of autonomic feedback based on the Mandler Autonomic Perception Questionnaire, Barron Ego Strength scale, Taylor Manifest Anxiety scale, and Nowlis Adjective Check List scores of 139 male and 37 female college students. Women are significantly higher in autonomic feedback scores, lower in ego strength and higher in manifest anxiety, though the within-sex correlations were substantially the same for both sexes. In both groups, autonomic perception correlated positively with manifest anxiety and negatively with ego strength; the latter two correlating negatively between

themselves. High APQ scorers had significantly higher scores on the Nowlis-Green factor variables of anxiety and depression. However, within each sex group, proneness to autonomic feedback may characterize those who are least successful in their sex roles. In adjective self-descriptions, high APQ males reveal themselves as ineffectual, dependent, and depressed, high APQ females as more active and aggressive with stronger and more labile emotions. Thus, it may be concluded that reporting more numerous and intense somatic experiences in emotional states is generally characteristic of more neurotic, inadequate, and anxious individuals, though the particular neurotic problem may differ for each sex.

Though more autonomic feedback is reported in anxiety than in happiness, the positive correlation between autonomic feedback scores in the two states suggests that the tendency for somatic involvement, at least as assessed by the APQ, is a general characteristic of individuals in a variety of emotional states.

REFERENCES

- BARRON, F. An ego-strength scale which predicts response to psychotherapy. *J. consult. Psychol.*, 1953, 17, 327-333.
- BARRON, F. A correction. *J. consult. Psychol.*, 1954, 18, 150.
- GREEN, R. F., & NOWLIS, V. A factor analytic study of the domain of mood with independent experimental validation of the factors. *Amer. Psychologist*, 1957, 12, 438. (Abstract)
- MANDLER, G., & KREMEN, I. Autonomic feedback: A correlational study. *J. Pers.*, 1958, 26, 388-399.
- MANDLER, G., MANDLER, JEAN M., & UVILLER, ELLEN T. Autonomic feedback: The perception of autonomic activity. *J. abnorm. soc. Psychol.*, 1958, 56, 367-373.
- NOWLIS, V. The development and modification of motivational systems in personality. In, *Nebraska symposium on motivation*, 1953. Lincoln, Nebraska: Univer. Nebraska Press, 1953. Pp. 114-138.
- NOWLIS, V., & GREEN, R. F. The experimental analysis of mood. Paper read at the fifteenth International Congress of Psychology, Brussels, July 1957.
- TAYLOR, JANET A. A personality scale of manifest anxiety. *J. abnorm. soc. Psychol.*, 1953, 48, 285-290.

(Received August 1, 1960)

THE BENDER GESTALT:

A CLINICAL STUDY OF CHILDREN'S RECORDS

WENTWORTH QUAST

University of Minnesota Medical Center

The need for indicators of brain dysfunction at all ages is apparent to clinical psychologists, but is particularly acute in the assessment of children where changes with age and wide variation of many growth characteristics within an age range complicate the problem. Incidence figures for brain injury in children in this country are highly variable, although approximately three million is an estimate made by Martha Elliott (1956), Chief of the Children's Bureau. This figure does not include an additional number of children with behavior and learning disorders presumed to result from organic pathology.

One of the main problems confronting the child clinical psychologist is in the relative weights to be assigned to emotional, functional, dynamic, or learned components as contrasted with intrinsic, constitutional, or organic components in the unusual symptomatology presented by the child patient. One measure of presumably intrinsic defects is difficulty in coordination as evidenced in visual-motor tasks. For research and clinical evidence that organic patients do have visual-motor difficulties, the reader is referred to reviews by Klebanoff (1945) and Klebanoff, Singer, and Wilensky (1954). The present study was undertaken to test the validity of the Bender as an indicator of organic components using 100 child subjects, either inpatients or outpatients of the Division of Child Psychiatry, University of Minnesota Medical Center.

The problem of establishing clear criteria for brain injury in children is complicated by the subjective nature of neurologic examination, by problems of reliability and validity in EEG interpretation, and by the frequent lack of any demonstrable physical changes in

a child with known central nervous system defect. A survey of the final diagnoses of 325 consecutive inpatient admissions showed acute or chronic brain syndromes to constitute about 25% of the population. Considering the lack of definitiveness in making such a diagnosis, this figure is probably low. In the search for criterion groups it was felt more meaningful to examine the original impressions of referring physicians or agencies, the presenting complaints, or the differential diagnoses considered at the time of admission to hospital for suspicions of brain injury. When these are considered and the problem becomes one of examining for possible central nervous system involvement, the proportion of suspected brain injured becomes more nearly 50% of the clinic population. If the problem of mental deficiency were to be included, the problem of determining "familial" versus "organic" etiology is also at least a 50-50 base rate problem. It should be clear that the terms "brain injury" and "suspected organic brain damage" are used in a categorical sense, to indicate a wide range of central nervous system defects. Also it should be clear, while inaccuracies on the Bender beyond a certain age may be related to cerebral defect, the extent of such defect is not implied. The test is examined as a screening device capable of separating out those children who warrant a "second look" neurologically.

METHOD

Since it was possible to identify some homogeneity among various presenting complaints of child psychiatric patients, criterion groups were designated according to whether or not brain damage was suspected: (a) those with suspected emotional disturbance without suspected brain damage and (b) those with suspected brain damage, with or without emo-

tional disturbance. Notations regarding any suspected organicity appeared in the hospital chart in the form of a physician's referral specifically for neurological study; in the form of behavioral complaints from physician or schools related by them to brain injury by inference, such as extreme hyperactivity, short attention span, impulsivity, etc.; and in the form of notes regarding questionable or known seizures, or notes as to sequelae from illnesses or trauma.

The emotionally disturbed group consisted of patients in whom brain injury was not mentioned as suspect by any person in the chain of individuals from the original source of referral through the final examining physicians in hospital. Presenting complaints common to this group were fears, obsessions, lying, stealing, truancy, pathological shyness or withdrawal, allergies, and somatic complaints. No child whose presenting complaint was mental retardation was included in either group.

The resulting two groups were matched relative to their socioeconomic level in proportions according to father's occupation using the Minnesota Scale of Parental Occupation. The lower age limit was 10 (9 years 10 months), and the upper age limit 12 (12 years 11 months). A limit of 1 year's age span would have been most desirable to insure homogeneity in this regard but sufficient numbers were not available within such a narrow limitation. Children aged 10 were chosen since normative data previously obtained by Quast (1957) indicated absence of most Bender

deviations by that age. A number of "organic signs" on the Bender seen even in adult records occur as normal developmental phenomena up to age 8 but usually not beyond 10. Thus, such an age selection controlled developmental deviation to large extent.

It was impossible to make the sex ratio comparable to the population at large since the usual clinic ratio averaged two boys to one girl. The final selection of patients was considered to be a representative sample of the clinic population.

The mean IQ for the emotionally disturbed group was 99.5, with a standard deviation of 15.72. Mean IQ for the suspected brain damage group was 81.7, with a standard deviation of 16.55. The difference in mean IQ was a reliable one with a *t* of 5.57, *p* = < .0001; however, previous data (Quast, 1957) would indicate IQ per se not to be a significant determinant in Bender performance. Also, the mean mental age of the brain damage suspects (9 years 1 month; *SD*, 2 years 1 month) exceeded the cutoff mental age 8, below which deviations occurred as normal developmental phenomena.

The two groups were similar in that they could be considered "typical" child psychiatric patients. They shared common referral sources and represented a wide range of problems.

The Benders were administered by the writer, or by advanced graduate students during their hospital internship in clinical psychology. All tests were scored according to the Peek-Quast system (1951)

TABLE 1
RELATION OF BENDER ATTRIBUTES TO CRITERION GROUPS (SUSPECTED BRAIN DAMAGED AND SUSPECTED EMOTIONALLY DISTURBED)

Attribute or sign	% Brain Damaged Suspects Showing Sign (N = 50)	% Emotional Suspects Showing Sign (N = 50)	Relation of Sign to Criterion Groups in Terms of ϕ	Level of Significance	% Normative ^a Ages 10 & 12 Showing Sign (N = 100)
Scalloping	24	2	.327	.01	1
Dashing	24	4	.288	.01	6
Perseveration	48	12	.393	.01	14
Rotations, 2 or more, +2 or +3, not Figures a, 3	32	2	.399	.01	6
Reversal	26	0	.387	.01	3
Confabulation	26	0	.387	.01	1
Angulation, +2	40	2	.483	.01	3
Mixed orientation	36	36	.000	—	11
Card turning, inversion	24	12	.156	—	0
Major distortion	56	4	.567	.01	7
Erasures, absence of	50	26	.247	.05	42
Excessive pressure	46	50	-.040	—	78
Separation, +2	20	0	.333	.01	6
Flattening	24	22	.024	—	28
Exaggeration	64	42	.220	.05	39
Line substitution	0	0	.000	—	0
Slope	66	34	.320	.01	38
Global clinical impression	80	12	.682	.01	

^a Presented for convenience in comparison.

TABLE 2
INTERCORRELATIONS BETWEEN ATTRIBUTES CHARACTERISTIC OF THE
SUSPECTED BRAIN DAMAGE SAMPLE

	2	3	4	5	6	7	8	9	10
1	-.20	+.43	-.06	+.06	.00	+.30	+.12	+.16	.00
2		+.19	-.19	.00	+.06	+.14	.00	.00	+.13
3			-.17	+.02	+.42	+.28	+.03	+.12	-.07
4				+.14	-.02	.00	.00	.00	+.04
5					+.27	+.10	+.18	+.30	+.25
6						+.38	+.57	+.25	+.12
7							+.56	+.24	+.25
8								+.30	+.04
9									+.39
10									

Note.—Correlations are phi coefficients. 1 = Scalloping, 2 = Dashing, 3 = Perseveration, 4 = Rotation, 5 = Reversal, 6 = Confabulation, 7 = Angulation, 8 = Major Distortion, 9 = Separation, 10 = Slope.

by the writer without knowledge of the patient's name or classification. A global judgment was made on appraisal of the record as a whole, on a no brain damage versus brain damage dichotomy, prior to scoring.

Analysis of the scored records consisted of comparing the two groups for the presence or absence of a number of Bender attributes or signs, found in the examiner's experience to be most common in the records of brain injured children and adults. An a priori selection was made of 17 attributes most likely to discriminate the two groups.

Item validity was obtained by the use of Jurgensen's tables (1947) for the determination of phi coefficients, and levels of significance were calculated. For those attributes showing the best discrimination (.01 level), intercorrelations were also calculated.

The scores were assumed to be dichotomous so the correlation method selected was the phi coefficient. A test of the null hypothesis can be made through phi's relationship to chi square. $\text{Chi square} = N\phi^2$. If chi square is significant in a four-fold table, the corresponding phi is significant at the same level. Specifically, where $N = 100$, chi square is significant at the .01 level if it is at least 6.6, and at the .05 level if it is at least 3.8.

RESULTS AND DISCUSSION

The results of the comparisons between the performances of the suspected brain damaged and the suspected emotionally disturbed groups are presented in Table 1 which expresses the relation of the Bender attributes to the criterion groups in terms of phi coefficients and the levels of significance. Normative data (Quast, 1957) on the 10- and 12-year olds combined are presented for comparison.

It is seen that 10 of the attributes separate

the two groups at the .01 level of significance. Two others showed significant differences between the groups at the .05 level. The fact that clinical impression of the record as a whole separated the groups best is not surprising when it is learned that the suspected brain damage group averaged 3.6 attributes per patient as contrasted with a mean of .60 for the suspected emotionally disturbed. False positive Bender signs in the suspected emotionally disturbed occurred to a significant extent on only 1 of the 10 discriminating attributes. This attribute, Slope, while occurring in two-thirds of the brain injury suspects, occurred in one-third of the suspected emotionally disturbed. The busy clinician should note that the most discriminating signs were generally also the easiest to score. These data suggest that a cutting score based on the signs which give optimal discrimination, possibly weighting the better signs, could be developed through a cross-validation study.

When intercorrelations were calculated for these 10 attributes (Table 2) it was found that in general they appeared to show little intercorrelation with one another. Because of interest in the underlying strength of relationship between signs rather than in making predictions from one sign to another, and recognizing the restriction which reduction of frequencies to 2×2 tables places upon phi coefficients, the obtained phi's should be interpreted for size in light of the maximal phi coefficients possible. About 8 of the 45 intercorrelations were appreciably affected by such

considerations (for example: $\phi_{13max} = .65$, when obtained $\phi = .43$; $\phi_{36max} = .61$, when obtained $\phi = .42$).

These data are consistent with the hypothesis that the Bender may be tapping a variety of defects which may exist, in combination or separately, in an individual patient. If this is true and each sign is valid, and the probability of their occurring in patterns is low (low correlation), then a count of critical signs takes on considerable significance for diagnostic purposes. Validity of a total score is much enhanced when separate parts or items are valid individually but modestly intercorrelated. However, a combinatory sign approach may be ultimately less helpful in terms of explanation of the deviations than a further exploration of a particular sign.

For example, while the attribute Perseveration can logically be linked with other known perseverative behaviors and thinking of the brain injured person, Scalloping, Rotations, Angulation, and/or others, may be linked with defects hitherto unexplored. Attention to individual signs may also throw light on the problem of whether a more fundamental, general disturbance (for example, "integration") may be involved which is expressing itself in a variety of ways, rather than associating signs with localized brain areas. That a more central, general, integrative phenomenon is operant is suggested in this sample by the findings that some visual motor disturbance was common to the group despite the wide variation of central nervous system complaints.

Seriously ill patients have at times been able to describe the reasons for their deviations, and several of their explanations have been illuminating. With some consistency patients in acute neurologic states have described a conservation of energy as the main reason for certain deviations. In verticalizing horizontally oriented figures, for example, patients have stated it "easier" to make flexor rather than extensor motions. Other acutely ill patients have described with some distress their inability to make an angle in one direction while demonstrating facility with angulation in another. That this difficulty in laterality has an appropriate cerebral morphological correlate or that defects in circumscribed cortical sensory or motor areas are at least "in-

strumental" prerequisites, would seem justified since peripheral muscle mechanics offer inadequate explanation in most cases.

SUMMARY

The validity of certain Bender deviations as indicators of cerebral dysfunction was examined in 100 child psychiatric patients, aged 10 to 12. On the basis of their presenting complaints, patients were divided into two groups designated as suspected brain damaged and as suspected emotionally disturbed. Mental age and socioeconomic status factors were controlled.

An a priori selection of 17 attributes normally not occurring after age 8 showed 10 of these to differentiate the two groups at the .01 level. False positive "organic" signs in the suspected emotionally disturbed occurred in but one discriminating attribute. Intercorrelations between the 10 attributes having the best discriminatory power showed them, in this sample, to have in general low positive correlations with one another.

These data suggest that the practicing clinician would do well, when these 10 attributes appear in the records of his patients, to consider "neuronic" rather than neurotic etiology for the child's behavior. For the researcher, it is hoped that these findings may serve as a building block toward spelling out the heterogeneous nature of that group now referred to categorically as brain damaged.

REFERENCES

- ELLIOT, MARTHA. Paper presented at Annual Meeting of Association for Aid to Crippled Children, May 1956.
- JURGENSEN, C. E. Table for determining phi coefficients. *Psychometrika*, 1947, 12, 17-29.
- KLEBANOFF, S. G. Psychological changes in organic brain lesions and ablations. *Psychol. Bull.*, 1945, 42, 585-623.
- KLEBANOFF, S. G., SINGER, J. L., & WILENSKY, H. Psychological consequences of brain lesions and ablations. *Psychol. Bull.*, 1954, 51, 1-41.
- PEEK, R. M., & QUAST, W. A scoring system for the Bender-Gestalt test. Minneapolis, Minn.: Authors, 1951.
- QUAST, W. Visual-motor performance in the reproduction of geometric figures as a developmental phenomenon in children. Unpublished doctoral dissertation, University of Minnesota, 1957.

(Received August 1, 1960)

A COMPARISON BETWEEN HYPNOTICALLY INDUCED AGE REGRESSIONS AND WAKING STORIES TO TAT CARDS:

A PRELIMINARY REPORT

JOSEPH REYHER

Michigan State University

AND

DONALD SHOEMAKER

Southern Illinois University

Estimating the degree to which a client in therapy has worked through important areas of conflict often imposes a difficult judgmental task upon the therapist. A method for increasing the objectivity of this task was developed in connection with an exploratory investigation involving the comparison of hypnotically induced age regressions to TAT cards as stimuli and subsequent waking stories to the same cards. The original intent of the study was to develop a procedure for reducing the artificiality of hypnotically induced conflict so the results of such data could be more meaningfully interpreted; however, the immediate diagnostic significance of the data overshadowed some of the more long range experimental goals. It was found that differences between the content of the hypnotic and waking reactions significantly extended diagnostic impressions, often reflected unresolved areas of conflict, were a useful psychotherapeutic aid, and provided valuable insights into hypnosis itself. Psychoanalytic theory served as the frame of reference for both the psychotherapy and the experimental design.

METHOD

Subjects

Five subjects were used. Because of the possibility that adverse posthypnotic reactions might occur as a result of the stimulation of a subject's emotional conflicts, only clients were used who were in or about to engage in psychotherapy. Furthermore, they were given some control over their emotional reactions in the waking state in order to reduce the possibility of premature activation of problems which involved severe conflict. Four of the five subjects were able to experience a complete posthypnotic amnesia. The one exception usually experienced only partial amnesia. Three of the subjects were characterized by predominantly neurotic reactions; the other two

were characterized by predominantly psychotic reactions, but they were sufficiently intact to function without hospitalization.

Procedure

Since the same TAT cards were not suitable for both sexes, two sets of 17 cards each were assembled.¹ Eight of the cards were common to both sets. A random procedure was used to determine, for each subject, the selection of 10 cards from the appropriate set, the designation of the cards as either conflictual (c) or neutral (n), and the order in which the cards were presented.²

Conflict cards. While hypnotized, the subject was told that he would be asked to look at a picture that would activate disturbing emotions. He then was asked to open his eyes and look at the card. After about 10 seconds, he was instructed to close his eyes and to go back in time to a period when these emotions were very difficult to manage. He was then asked to verbalize his experience.

Neutral cards. The instructions were the same as for the c-cards except that the subject was told that the emotions to be experienced would not be disturbing but, nevertheless, would be meaningful.

Posthypnotic suggestions. All subjects were given the following instructions:

Sometime later, when you are awake, Dr. A. will give you the same pictures that I gave you earlier, and he will ask you to tell stories about them. Each picture will stir up the same feelings, emotions, and ideas that it did before, but you will

¹ The cards unique to the female set were 3GF, 18GF, 13G, 12F, 9GF, 7GF, 8GF, 17GF, and 2; the cards unique to the male set were 8BM, 18BM, 20, 14, 6BM, 7BM, 12BM, 13B, and 1; the cards which were common to both sets were 5, 10, 4, 3BM, 12BG, 6GF, 13MF, and 17BM. Cards 15, 16, 11, 9BM, and 19 were not used.

² There was one restriction on the random procedure which was used for determining the conflictual and neutral designations of the cards. The structure of card 13MF was too extreme in a conflictual direction to risk giving it a neutral designation.

either reveal them directly or indirectly in the stories that you tell.

The last sentence in these instructions was intended to give the subject control over the nature of what would be consciously experienced in order to reduce the possibility of a traumatic reaction to the premature recognition of repressed material.

Waking condition. The subjects were given the cards by Dr. A. with standard instructions.

RESULTS

All subjects produced stories in the waking state that varied in the extent to which they responded with the hypnotic reactions to the same cards. In order to evaluate objectively the degree of similarity between the two sets of data, the reactions to each card for the two conditions were compared in terms of three of the most obvious dimensions of similarity which repression may influence: characters, situations, and affective-motivational state. The degree of similarity for characters (C) and situations (S) was assessed by a four-point rating scale with the following descriptive labels and numerical values: personalized (O), congruent (1), indeterminant (2), and different (3). The affective-motivational state (AM) was quantified by counting the expressions of intentions, needs, and affects by the subject in the hypnotic condition (H) and by the corresponding character in the post-hypnotic condition (PH). The AM units in the latter condition were classified as either the same (PH_s) or as different (PH_d) from the hypnotic condition. Each AM unit was counted only once in each of the two conditions, regardless of how often it may have appeared.

A difference score (D) for each card was devised:

$$D = C + S + AM$$

where:

$$AM = \Sigma H - PH_s + 2\Sigma PH_d$$

If no changes occur in the waking reactions to a card, D equals zero. As changes occur in the waking state, PH_s decreases and PH_d increases. The term PH_d is weighted by a factor of 2 because it represents a transformation of affect, which is considered to be a more complicated process than simple repression.

The D scores for the five subjects ranged from 5-16, 4-15, 4-16, 6-13, and 5-23. Low

and high scores were arbitrarily defined as those which were in the first and fourth quartiles, respectively.

An estimate of the interrater reliability for two raters was obtained from rank order coefficients of correlation between D scores for two of the protocols that were selected randomly. Coefficients of .74 and .75 were obtained and were considered to be sufficiently large for research purposes.

The following protocols were selected in order to illustrate the type of material that was elicited by the c- and n-cards which had high D scores. Subjects A and B were similar insofar as initial psychodynamic impressions indicated that they both had unresolved Oedipal conflicts, nonspecific sexual fantasies concerning an older brother, and confused psychosexual identifications which appeared to be related to parental preference for an older brother.

Subject A, Card 4 (c-Instructions)

Hypnotic reactions. He reminds me so much of daddy. I don't know—he just seems always to have some barrier—something between us. So many times I'd like to tell him I love him or give him a big kiss or something, but I can't. I'm afraid he wouldn't—I don't know—I don't think he would push me aside or anything but I just never have told him how I liked him for a father and all the things he's done for me and I can't get close to him and I'd like to be able to but he just—I don't know why—maybe he's afraid of women, but I want to so bad all through my life. When I was little, I would sit on his knee and play with his hands—'cause they were bigger than mine, and I had so much fun that way. When I got older and wanted him to still treat me like that or show that he still liked me, he just wouldn't, and I almost hated him for it—'cause other girls' dads kissed their daughters and took an interest in them. It seemed he just let me go my own way and let mother take care of me and I didn't want her to—I wanted him to. He just wouldn't, but yet I would—I do things for him. Like I'd go golfing with him, and I'd drink some beer with him during the summer sometimes on a hot night—to please him, but that's all I did it for was to please him, but I liked it. But I can't—it seems like—I wish I could.

Waking reactions. Oh, ah-h, let's say, I believe they are married—these two people—and he's the kind of a person that would love to tell his wife how he feels about her but he can't quite and she almost craves it. She wants him to tell her all those little sweet nothings and he finds it hard to do so and therefore—they don't understand one another and she—tries to console him, but she can't because she doesn't understand it.

These reactions reinforced the initial psychodynamic impressions concerning sexual inclinations toward her father.

Subject A, Card 8GF (n-Instructions)

Hypnotic reactions. I'm watching my brother and he's painting, and I admire him so much. I think he's kind of like a genius or something—the way he can create pictures out of paint. I couldn't do that and I'm kind of jealous of him. Still, he was my brother and I thought he would make something of himself and *I could be his sister*. He used to make the most beautiful scenes. He couldn't paint people, just scenery and things. They were so pretty.

Waking reactions. This girl really has a dreamy look in her eyes. She is in a classroom and she admires her—oh—English teacher very much. She thinks to herself that someday she will be just exactly like *her*. Is that short enough? That is, I mean, does it make any difference how long or short they are?

Her underlying attraction to her brother becomes very clear. Substitution of the word "wife" for her illogical use of "sister" would be more congruent with unacceptable fantasies concerning her attraction to her brother. At no previous time had there been any evidence that homosexual tendencies might be involved. The waking reactions indicated that these tendencies may be in the service of defense against incestuous fantasies. The obvious defensive disturbances in the waking reactions reinforce these impressions.

Subject B, Card 7GF (c-Instructions)

Hypnotic reactions. I am thinking about two things and I can't separate them. (What are they?) It seems that I'm sitting on my mother's lap and she is telling me about having my first menstrual period and I also seem to think at the same time about being tied in a chair. I can't—I can't quite remember, I know I used to get tied in the chair when I was bad and I don't know why.

Waking reactions. The mother is reading a story to the little girl—and the little girl is holding a doll in her arms—and tells her that someday she might become a mother—and be the kind of a person that her mother is reading about in the book. (How is she feeling?) She feels as if she would like to grow up and it seems a long way off. (How will it end for her?) She will grow up and have a baby. (Laughs.)

In subsequent therapeutic sessions, successive dreams were induced in the subject regarding the emotions and thoughts behind the image of sitting on her mother's lap learning about menstruation and the image of being tied to a chair. The induced dreams to men-

struation were related to psychosexual confusion which progressed to an abreaction of her anxiety regarding her wish to be a boy. The image of being tied to a chair was related to a traumatic early childhood experience in which she heard a voice, while she was in bed, telling her to kill her mother.

Subject B, Card 17GF (n-Instructions)

Hypnotic reactions. My daddy and I used to ride on bicycles through the—part of the canal—and it was very enjoyable for me—by the water and a bridge. I always enjoyed it—riding on a bicycle there with my dad, because we always liked the water—threw stones in it. (Is there anything else?) I—used to like it, that's all.

Waking reactions. It seems to be a granary of some sort—there are sacks of grain on the ground. There's a girl on a bridge. In the background is the granary and she would like to leave it, but she can't make up her mind whether she wants to or not. I think she would like to be on a boat and go away. (What will finally happen?) I don't think she will. (Can you make up a story about why she is torn between leaving and staying?) She feels that the people in the house will miss her if she goes away. She would like to go somewhere else.

The material produced in the hypnotic reaction was related to the many dreams she had concerning water. Successive dream induction produced the reliving of an experience in which she was riding on her father's back in the water while having thoughts that she wanted to possess him completely.

Subject C was a tense, conscientious, married female who was successfully engaged in one of the medical specialties. Marital difficulties constituted the presenting problem.

Subject C, Card 17GF (c-Instructions)

Hypnotic reactions. This is my fear of high places for myself or to see someone else in a high place, even in the movies or TV, when they're out on a ledge. When we were at the Grand Canyon, other people would lie on their stomachs and would look over the edge. Why, I don't know, unless heights symbolized to me jumping and suicide—trying to stay away from the situation or avoid it. Well, it's not so much fear of jumping myself as an accident, of someone *not meaning* to but *accidentally pushing* someone over or falling. *Well intentions* and yet *unavoidable*. I have thought briefly in terms of suicide but whether I would ever be able to commit suicide, I decided that I would never be able to. (Anything else?) No, I don't believe so.

Waking reactions. A woman on a high bridge feeling very despairing, uh—considering jumping. She feels that things haven't gone as they should but she isn't going to jump.

The italicized portions reveal the underlying hostile impulses which were betrayed in terms of rather direct verbal representations. The attempts to cover up were ineffectual, and defensive reactions due to the marked breakthrough of murderous impulses were not present. Any such slips in the waking state probably would have activated a host of autonomic reactions and vigorous defensive activity. The brief and impoverished waking reaction supports this interpretation. No high D scores were associated with n-cards.

DISCUSSION

In view of the omnipresent possibility of artifact due to the motivation of hypnotized subjects to please the experimenter by behaving in a manner consistent with what they believe is expected of them, caution must be exercised in the interpretation of the results. In terms of the procedure, it would seem likely that subjects would perceive that the instructions to the c-cards were of special interest. If subjects had surmised correctly that the experimenter had expected greater differences between the hypnotic and waking reactions to c-cards than to n-cards, there should have been higher D scores associated with the former. Although this was true, the material that was produced was more closely involved with potent areas of conflict than material which had come up at that point in therapy. Aside from whatever role a desire to please may have played in producing the material, it proved to be extremely useful in achieving a more sharply defined and delineated psychodynamic impression.

The enhanced clarity of the psychodynamics to c-cards with high D scores supports White's (1941) observation that hypnosis is an altered state of awareness in which there is a contracted frame of reference and lack of such critical functions as a sense of humor and self-consciousness. It may be that a reduction in these functions proportionately reduces the anxiety producing potential of repressed intrapsychic stimuli which enables them to become more clearly represented in awareness. The most perplexing and unexpected data, however, were the high D scores for a number of the n-cards to which the subject had experienced positive affect in the hypnotic reactions. Previous psychodynamic im-

pressions indicated that this was not a case of simple forgetting, because the high D scores involved material related to the most central areas of conflict. Unlike the c-cards, material of equal or greater anxiety producing potential was reacted to with positive affect. The two sets of instructions appeared to elicit different kinds of unconscious processes or different facets of the subjects' conflicts.

The observed differences in the reactions to c- and n-cards with high D scores are probably related to the difference in the instructions for these cards. The c-instructions required the subject to react with disturbing affect to whatever was brought to mind by the card. In response to this, the subject seemed to recall some previous conscious experience that was distressing to him. When this reaction was experienced, underlying repressed aspects of this material became more clearly represented. The high D score, therefore, seems to be a function of the potential threat of perceiving this same material with the enhanced critical functions of the waking state.

The n-instructions merely required the subject to respond with reactions that were not disturbing but, nevertheless, meaningful. Unlike the c-instructions, the n-instructions made a comparatively nonspecific request, insofar as the subjects' reactions to the card were not restricted to one class of affect. It may be that the positive reactions are a function of hypnosis as an altered state of awareness, and that the same material would activate anxiety in the waking state. This interpretation is consistent with psychoanalytic theory in that anxiety and defense are produced by repressed material that is striving for an outlet and gratification. Again, the high D score probably is a function of the enhanced critical abilities of the waking state in which the perception of such material would activate a prohibitive degree of anxiety.

In order to pursue the implications of the above interpretation of the high D scores to n-cards, we have been instructing clients under hypnosis to let their minds wander as they do just prior to dropping off into a deep sleep, with images, dream-like thoughts, and feelings coming without any conscious effort. The client is not instructed to talk until he is awakened, at which time he is simply asked to talk about his reactions. The suspension of

conscious effort and talking are designed to reduce critical functions (secondary process) that would be activated by the need to communicate. Although our experience with this method is still limited, the results have been very encouraging. When the spontaneous images, fantasies, dreams, and memories which are produced become too transparent, a defensive posthypnotic amnesia may occur; however, subsequent hypnosis frequently can recover much of this material.

Cards 8GF and 7GF for Subjects A and B, respectively, illustrate that alterations of affect and motivation can masquerade as health. Most of the waking reactions, however, contained some clues as to the presence of underlying conflicts, but the alterations often were so great that the nature of the actual conflict could not be readily inferred without benefit of the hypnotic reactions. The most productive method of analysis with these data was to consider the hypnotic and waking reactions jointly. The hypnotic reactions were often closer to conflictual material, whereas the waking stories gave a better picture of defensive reactions. Thus, successive dream induction and a combined analysis of both types of TAT data present a better opportunity to assess the nature of the underlying anxiety producing processes and attendant defensive reactions at three different levels of psychic representation.

Not only do high D scores provide valuable diagnostic information about the status of the subject's conflicts and the degree to which he has worked them through, but they sometimes open up areas previously unknown to the therapist. One method of utilizing the high D scores in therapy is to give the subject the posthypnotic suggestion to recall all the age regressions, with the provision that he can forget what he does not want to remember at this time. If the age regressions to the cards with high D scores are not recalled by the subject in the waking state, further evidence of the dynamic significance of the material is obtained. These areas then become the foci of subsequent hypnoanalytic sessions.

It is not surprising that the subject may remember some of these age regressions in the waking state because the hypnotic material is still relatively disguised. If subsequent waking and hypnotic interviews are unproductive in

further uncovering and working through of the material denoted by a given high D score, specialized hypnoanalytic procedures can be used. One such procedure which is particularly well suited for this purpose is successive dream induction concerning the emotions and experiences that produced the content of the age regression. Examples 7GF and 17GF for Subject B illustrate this procedure.

SUMMARY

The clinical and experimental value of hypnotic and waking reactions to TAT cards was investigated. Ten TAT cards were selected randomly to be conflictual or neutral for five psychotherapy clients who were capable of deep hypnosis. The conflict-inducing instructions were intended to activate disturbing reactions to TAT cards. This was followed by an age regression to a period earlier in life when the activated reactions were particularly difficult to manage. The nonconflict-inducing instructions were the same except that they were intended to activate nondisturbing, but meaningful, reactions.

A difference score (D score) was devised in order to evaluate quantitatively differences in characters, situations, and affective-motivational states between the hypnotic and waking reactions to each card. Both conflict cards and neutral cards had high D scores which reflected areas of conflict that had not yet come up in therapy and areas that had not been worked through adequately.

Neutral cards with high D scores, unlike conflict cards with high D scores, were associated with positive affect. It was concluded that the conflict cards activated previous conscious, conflictual experiences which more clearly revealed underlying repressed material than the waking reactions to the same cards. The high D scores to the neutral cards were interpreted as evidence that hypnosis is an altered state of awareness in which unconscious drives tend to be perceived in terms of gratification rather than threat. A procedure was described for utilizing this material in psychotherapy.

REFERENCE

- WHITE, R. W. A preface to a theory of hypnotism. *J. abnorm. soc. Psychol.*, 1941, 36, 477-505.

(Received August 3, 1960)

ANXIETY IN VERBAL BEHAVIOR:

A VALIDATION STUDY¹

MERTON S. KRAUSE

Family Service of Cincinnati

AND

MARC PILISUK

University of Michigan

Several investigators have suggested that the degree to which a person's speech departs from its usual level of coherence and economy is a likely indicator of anxiety (Dibner, 1956; Eldred & Price, 1958; Mahl, 1956). The following experiment was designed to evaluate the predictive validity of measures of such speech disruption as indicators of anxiety.

MEASURES

Both Mahl (1956) and Dibner (1956) describe measures of speech disruption. The two measures appear similar and produce highly correlated total scores: the median for 15 cases was .91 (Krause, 1961a). We felt that these measures could be improved with regard to their "psychological meaningfulness," reliability, and range of sensitivity. In the first place there was a good deal of overlap between the two measures. Mahl's "sentence incompleteness," "repetition," and "stutter" (plus "omission") categories resemble Dibner's "unfinished sentence," "repeating words or phrases," and "stuttering or unfinished words," respectively. We attempted to regroup the various speech disruptions into categories which referred more uniformly to the expression of complete thoughts and which might be interpreted in terms of the function served by the disruption in the flow of speech. To these ends we adopted Dibner's "break," Mahl's "correction," their common "unfinished" or "incomplete," and pooled their "repetition" and "stutter" categories. Mahl's

"tongue slip" and "omission" and Dibner's "omission" (part of his "stutter" category) we combined into "distortion," Dibner's "I don't know," "sigh," "laugh," and "question," as well as Mahl's "intruding incoherent sound," were pooled in our "intrusion" category. The Mahl "Ah" category we expanded as "procrastination." The following set of descriptions specifies our categories of speech disruption.

SPEECH DISRUPTIONS

B (break). The continuity of one line of thought is broken by the intrusion of another. The break is usually grammatically disruptive but it is the interruption of an incompleting line of thought by another, different thought that is distinctive. There must be an actual shift of topic, so that the interrupted material seems to have been displaced from the speaker's attention. Therefore a B is not scored when the interjected material is a commentary (even if rather anticipatory) upon or a correction of the interrupted material. Thrown-in statements, references to the interviewer, and questions can be breaks.

C (correction). Something is stated and then corrected within the same sentence. It may be a word, phrase, or clause which the correction is to replace, but in all cases the subject of the corrected and correcting material seems to be the same. Thus, the correction is usually one of pronunciation, emphasis, grammar, form, degree, specificity, or fact. It must not modify nor explain but *replace* the corrected material. The subject's conception seems to be that he has said the wrong thing, has misspoken himself, and must replace his word or words. A clue to this is that both

¹ This research was supported under a grant, M-516 C-7, from the National Institute of Mental Health, United States Public Health Service, E. S. Bordin, Principal Investigator. The authors are indebted to Joy Collins, Phyllis Pilisuk, and especially, Martin Timin for their assistance.

corrected and correcting elements fit the context.

F (fragment). A clause or sentence is left incomplete either in meaning or grammar. Whenever such a fragment occurs it is scored an F, even if it is apparently completely expressed in another independent try. Thus, when a sentence is interrupted by a qualifying clause and then the interrupted portion is repeated or replaced, this portion is a fragment regardless of how similar are the original and resumptive materials. An intended clause interrupted and replaced by another is an F, although the larger sentence into which both the interrupted and interrupting clauses fit is not an F. Answers to questions must be judged as to whether they are adequately complete responses so not judged independently of the question. When the material following a possible F could be construed to complete it, continuity or discontinuity of intonation is a clue, for an F often involves a change between the fragment and what follows it.

D (distortion). Mistakes or distortions of proper speech occur involving either grammar or meaning. They may be unintended words, including neologisms and tongue slips, incorrect words or grammatical errors which are not habitual, or mispronunciations. These distortions are recognized by their improbability, *prima facie* or as inferred from the context (often from the correcting material). Word omissions are scored Ds where without the ostensibly omitted word neither portion of the statement (before or after the omission) would stand as a complete sentence or thought.

R (repetition). Certain phrases, words, or parts of words are perseverated upon by repetition rather than prolongation. The flow of speech is impeded by redundant elements such as stutters, word or phrase repetitions, and changes between contracted and uncontracted forms. Repetitions share certain characteristics with the repeated material which distinguish their perseverative quality. These characteristics may involve pitch, timbre, loudness. The paradigm for R is the stutter.

P (procrastination). The speaker seems to procrastinate, to delay getting on to his next point. He does this by means of delaying sounds (such as "ah," "um," or prolonged

vowels), words (such as "well"), or phrases ("How shall I say?" sometimes even "I don't know"). That he is controlling his speech and thinking ahead distinguishes procrastinations conceptually from repetitions. Many phrases which also serve other functions may partake of procrastination. These include "let's say, let's see, as a matter of fact, that is, for example, so to speak, I mean," etc. "Well" and "Oh" are to be scored as procrastinations unless they clearly lack this quality, as when they are purely exclamatory.

I (intrusion). A nonverbal sound intrudes upon the flow of speech, but it occurs meaningfully like a break, rather than as a mere supporting procrastinating or background noise. It may be a sigh, laugh, cough, throat clearing, deep breath.

The reliability attained by two independent raters on the experimental material, rating from typescripts with the recording being played, was 88%. This represents exact agreement, i.e., the number of disruptions scored identically as to location and category divided by total disruptions scored. Contained in the denominator are the number of disruptions scored by one and not the other rater, scored differently by each, and scored identically by each. Over our protocols this coefficient ranged from a low of 71% to a high of 93%, with an interquartile range of about five percentage points. One-fourth of the protocols were rated after a lapse of 1 month, but the interrater agreement did not diminish. In order that response scores might be comparable over individuals with different speech outputs, scores were defined as the number of instances of occurrence for a particular category divided by the total number of words in the response.

EXPERIMENTAL DESIGN

The crucial design problem in this study was to provide anxiety criteria against which to test our measures. We employed a two-part criterion for recognizing instances of anxiety and nonanxiety. First, we developed a set of descriptions of what seemed stressful and non-stressful situations (i.e., stimuli). Second, we questioned the subjects as to what their feelings were while they were describing their probable reactions in these situations. Those

verbal responses to the presumably stressful situations which were reported to involve "a major element of tension, uneasiness, or apprehension . . . fear or fright" or anxiety, and no other emotions, were used as criterion instances of anxiety; while the responses to presumably nonstressful situations which were reported to involve no such element or any other emotion, were used as criterion instances of nonanxiety. We set aside, as irrelevant for validation purposes, the data on responses in which a stressor invoked no report of a fear or anxiety feeling, a nonstressor did invoke such a report or either invoked the report of some other emotion.

These criteria were used because they represented, in our opinion, the best evidence of anxiety that could be collected under the circumstances. The argument upon which this opinion rests is presented in detail elsewhere (Krause, 1961b), also see Grinker, Korchin, Basowitz, Hamburg, Sabshin, Persky, Chevalier, and Board (1956). Briefly, certain stimuli are on the face of them more stressful than others. Our confidence that anxiety is produced by them is greater than that it is produced by apparently innocuous stimuli. Our rationale for using the introspective reports, the second part of our criterion, is that one cannot feel afraid or anxious and yet not *be* subject to this emotion, for the feelings are *sufficient* evidence of the emotion's presence. It is possible, however, for one to *be* anxious without feeling anxious, thus if a nonstressor which induces no *feeling* of anxiety does produce speech disruption, the effects of predictive invalidity and "unconscious anxiety" will be confounded. The difficulty this involves for us is that we may get considerable speech disruption on trials classified as nonanxiety trials. This would tend to diminish the difference in speech disruption between these control trials and the experimental (anxiety) trials, thus decreasing the sensitivity of our experiment.

The experiment was performed, ostensibly, as a survey of how people react in disasters, under the title Disaster Relief Research. The subjects were recruited by means of a letter aimed at arousing both their interest and public spirit. It was sent to a systematic random sample (Cochran, 1953, pp. 160-168) of

University of Michigan students drawn from the summer directory. Only 19 of the 60 recipients volunteered to act as subjects, but we have no reason to believe our results biased by the subjects' self-selection.

Reinforcing the impression created by the letter, an introduction was given to each subject before the experimental session. The study was presented as an attempt to discover what public reaction to disasters would be. We stressed the practical value of such knowledge, its scarcity, and the subject's ability to provide it. The interviewer endeavored to make the session a collaboration in which the subject did his best to place himself in the situations described and spontaneously report his reactions, while the interviewer asked for elaborations or clarifications. This "cover" was essential to promote the subject's involvement, without which the stressors would be ineffective, and his honesty in reporting feelings, without which we could *not* assert any correspondence between feelings and reported feelings. The control achieved by our cover, like any control upon which it is infeasible to collect independent data as to its efficacy, is open to question, of course, but we have no reason to suspect that the subjects did not believe and report as we intended. Their involvement is evident in the recordings of their sessions.

The greatest danger of subjects' dishonesty biasing our results is in their possible tendency to report conventional, socially acceptable feelings. We were especially concerned about reports of anxiety feelings' presence or absence in response to what the subjects supposed the interviewer would expect, rather than to the subjects' actual feelings. The situations were not ambiguous within pairs as to the relative degree of danger, discomfort, or emotional arousal supposed to be involved.²

² The existence of such a social acceptability response-set is questionable here, since 58 out of 190 reports of feelings (i.e., from 1 to 9 for each subject) were unconventional in this sense.

One collateral finding is interesting in this regard. The unconventional responses of the 11 females in our sample show a surprising regularity. Those that scored over 12 on a 50-item sample of the Manifest Anxiety scale err only in reporting anxiety to nonstressors, while those scoring under 12 err only in reporting nonanxiety to stressors. The degree of this

This would tend to give us conservative results in the predictability of anxiety by speech disruption, i.e., any positive findings will tend to be understated.

We selected 10 pairs of situations, a stressor and a nonstressor in each. They all had some relevance to disasters and ranged from a first aid training proposal and Conelrad, through frustrated attempts to help others, to situations of great personal danger. They were paired in an attempt to minimize the differences, other than stress, between each stressor and its nonstressor mate. Thus, their positions in the series of 20 were contiguous and the degree and manner in which the response was structured by the situation and probe questions were equated within pairs. A greater similarity of within pair subject matter would have been desirable as well, but for the greater danger of overlap between the emotional effects of stressor and nonstressor. The interpair order was arranged so that the subjects would not perceive a stressor-nonstressor pattern and so that the apparently more intense stressors did not have a carryover effect on other stressors.

In summary: We exposed each of 19 subjects to 10 pairs of stimuli presented in a standard order. The subjects were under the impression that they were participating in a survey of how people might be expected to feel and react during disasters. Those of their pairs of responses which met our validity criteria were studied in terms of a measure of speech disruption for the relationship between anxiety and this measure.

RESULTS

Thirty-seven pairs of responses over 13 subjects (from a total of 190 pairs over 19 subjects) were appropriate for study. We were relatively ineffective in controlling the stressfulness of our stimuli, because we lacked prior knowledge of the subjects and used a standard set of stimuli for all subjects. Some recent work of Lazarus (1960) may prove helpful in the future selection of stimuli.

The seven categories of speech disruption were individually scored for each of the 37 "error" varies positively with MA scale score, $r = 0.77$ (significant at the .01 level).

TABLE 1
SPEECH DISRUPTION CATEGORIES AND PERCENT CORRECT, INCORRECT, AND INDETERMINATE

Prediction	Category						
	I	F	R	P	C	D	B
Correct	78*	59	54	51	43	24	5
Incorrect	14	38	43	46	38	11	0
Indeterminate	8	3	3	3	19	65	95

* Significant at the .05 level, one-tailed.

response pairs. If the sign of the difference (anxiety response score minus nonanxiety response score) was positive the category predicted correctly for that pair. If the sign of the difference was negative, it predicted incorrectly. And if there was no difference, it failed to discriminate. The differences for each category, alone and in combination with others, were tested for significance by the normal approximation to the binomial for the difference between two percentages. The results in Table 1 are the percentages obtained for each of seven categories viewed separately.

By combining categories we can obtain a slightly higher percentage of correct prediction than afforded by any individual category alone. The best simple combination seems to be to use I, and where it fails to discriminate use F. This combination predicted correctly in 86% of the cases and incorrectly in 14% of the cases.

Both Mahl and Dibner have used the sum of several categories for a speech disruption score. We can approximate their scores by combinations of our category scores. The Mahl non-Ah ratio (Mahl, 1956) is akin to the sum of all our categories excluding I and P, while Dibner's Cue-Count 1 (Dibner, 1956) resembles our sum minus only I.

Despite the intended similarity between the Mahl non-Ah ratio and our sum of categories excluding I and P, it should be pointed out that our approximation of this ratio discriminated correctly in only 57% of the cases. A somewhat higher figure was reported by Mahl and Kasl (1958), in a similar experiment, for the non-Ah ratio suggesting the possibility that the scoring criteria were not identical.

TABLE 2

SOME COMBINATIONS OF SPEECH DISRUPTION CATEGORIES AND PERCENT CORRECT, AND INCORRECT, INDETERMINATE

Prediction	Combination		
	Total	Non-I	Non-I & Non-P
Correct	70*	62	57
Incorrect	30	38	40
Indeterminate	0	0	3

* Significant at the .05 level, one tailed.

It does not seem that the extra effort required to score the categories other than I (with the possible exception of F) is warranted by their incremental effect upon validity. It is interesting, however, that speech disruption categories as a whole are, when pooled, statistically significant indicators of anxiety. This suggests that speech disruption may be a useful (unitary) concept.

In order to give the reader some basis for comparing the validities of the disruption measures with those of other verbal anxiety measures, we include the results on three other suggested measures for which it was feasible to score our data. These are number of words spoken in the subject's response (Balkan & Masserman, 1940), the verb-adjective ratio (Balkan & Masserman, 1940), and the latency of the subject's response (Benton, Hartman, & Sarason, 1955). The results, portrayed in Table 3, emphasize the effectiveness of I (with or without F) as an indicator of anxiety.

TABLE 3

NONSPEECH DISRUPTION CATEGORIES AND PERCENT CORRECT, INCORRECT, AND INDETERMINATE

Prediction	Category		
	Number of Words	Verbs/Adjectives	Latency
Correct	59	65	62
Incorrect	41	35	24
Indeterminate	0	0	14

DISCUSSION

The sorts of explanations alternative to anxiety which may be offered to account for speech disruption depend upon the situation in which the subject is placed. If we ask him to describe vague, ambiguous, or half forgotten experiences, then procrastinations, corrections, and fragmented sentences may well be frequent. If the interviewer is rather reactively mobile in his facial expression, this may induce a good deal of repetition or correction by the subject. An overly friendly or humorous interviewer can multiply intrusions of laughter, while one who nods his understanding anticipatorily may encourage distortions and fragmentation. If the subject does not perceive the situation as one calling for particularly coherent, grammatical, task-oriented, and economical speech—as he might not if he did not consider his role a serious one or if the interviewer were an old friend—he may not try to achieve a high level of disruption-free speech. We attempted to avoid these sources of speech disruption by our choice of stimuli, instruction of the interviewer, and experimental cover. Furthermore, these influences tend to affect the absolute level of speech disruption and not its variations over the several stimuli presented.

There are, however, peculiarities of our experimental situation which may limit the generalizability of our major results. The subjects were asked, in effect, to play-act. There were no "real" dangers in the situation perceptible by the subjects but for becoming excessively distraught by playing too well or offending the interviewer by getting too involved in the play and thereby "revealing" themselves. These dangers might be relieved by maintaining or shifting into a role-alien (e.g., an observing ego) viewpoint and so achieving some perspective or "emotional distance" from what one was saying or thinking. The predominant component of I, laughter, was, as we defined it, peculiarly well fitted as an indicator of such a stratagem. Where the subject became threatened, he may have laughed as an indication of his lack of commitment to the ongoing remarks, for one can most certainly discern this subtle intrusive quality in much of the subjects' laughter. This is not to say that

the subjects were not anxious then, but it does suggest that in a more thoroughly real or spontaneous situation (e.g., using sudden intense stimuli as stressors) intrusive laughter might be more rare, as Bs were in this situation, and so fail as a predictor of anxiety.

SUMMARY

In order to assess the predictive validity of speech disruption as an indicator of transitory anxiety, we exposed 19 subjects to 10 stressors and 10 nonstressors. The combination of stressor and reported feelings of anxiety was the criterion of anxiety's presence and that of nonstressor and reported absence of anxiety feelings was the criterion of anxiety's absence. We found that intrusive non-verbal sounds, mainly laughs and sighs, were the most correct predictors.

REFERENCES

- BALKAN, EVA R., & MASSERMAN, J. H. The language of phantasy: III. The language of patients with conversion hysteria, anxiety state, and obsessive-compulsive neurosis. *J. Psychol.*, 1940, 10, 75-86.
- BENTON, A. L., HARTMAN, C. H., & SARASON, J. G. Some relations between speech behavior and anxiety level. *J. abnorm. soc. Psychol.*, 1955, 51, 295-297.
- COCHRAN, W. G. *Sampling techniques*. New York: Wiley, 1953.
- DIBNER, A. S. Cue-counting: A measure of anxiety in interviews. *J. consult. Psychol.*, 1956, 20, 475-478.
- ELDRED, S. H., & PRICE, D. B. A linguistic evaluation of feeling states in psychotherapy. *Psychiatry*, 1958, 21, 115-121.
- GRINKER, R. R., KORCHIN, S. J., BASOWITZ, H., HAMBURG, D. A., SABSHIN, M., PERSKY, H., CHEVALIER, J. A., & BOARD, F. A. A theoretical and experimental approach to problems of anxiety. *AMA Arch. Neurol. Psychiat.*, 1956, 76, 420-431.
- KRAUSE, M. S. Anxiety in verbal behavior: An intercorrelational study. *J. consult. Psychol.*, 1961, 25, 272. (a)
- KRAUSE, M. S. The measurement of transitory anxiety. *Psychol. Rev.*, 1961, 68, 178-189. (b)
- LAZARUS, R. S. A program of research on psychological stress. In J. G. Peatman & E. L. Hartley (Eds.), *Festschrift for Gardner Murphy*. New York: Harper, 1960.
- MAHL, G. F. Disturbances and silences in the patients' speech in psychotherapy. *J. abnorm. soc. Psychol.*, 1956, 53, 1-15.
- MAHL, G. F., & KASL, S. V. Experimentally induced anxiety and speech disturbances. *Amer. Psychologist*, 1958, 13, 349.

(Received August 4, 1960)

CHARACTERISTICS OF TERMINATORS AND REMAINERS IN CHILD GUIDANCE TREATMENT

ALAN O. ROSS AND HARVEY M. LACEY

Pittsburgh Child Guidance Center

Caseload statistics at Pittsburgh Child Guidance Center indicate that 28% of all families accepted for treatment terminate their contact unilaterally and at a time which the clinic staff considers premature. This experience seems similar to that of other clinics, for Levitt (1958) states that some report a drop-out rate of more than 30%. Premature termination is costly because as many as 15 hours of scarce professional time may have been spent in diagnostic study and related activities before the family decides to discontinue the contact. This poses a practical problem. At the same time these families are of theoretical interest because, though carefully selected, they are unable to make use of collaborative treatment, the traditional approach of child guidance clinics.

Some families seem to derive benefit even from a truncated contact (Inman, 1956) so that the professional time may not be wasted completely, but if potential terminators could be identified early in the proceedings, they might be provided with a modified service, more suited to their needs. The staff time thus freed could then be made available to other families better able to use the traditional child guidance approach.

This study was designed to explore variables which might help differentiate between terminators and remainers. A number of investigators have attempted to identify characteristics of families who remain in treatment and of those who terminate prematurely. Levitt (1958) found that judgments of clinicians, made on the basis of diagnostic information, did not identify "defectors" successfully; nor did judgments of motivation

for treatment and severity of symptoms differentiate between the two groups.

Hofstein (1957) has pointed out that in a child guidance clinic, assessment of a child's treatability must rest on the evaluation of the parents' capacity to involve themselves in the treatment process and to work toward a change in their relations to each other as well as to their child. The important contribution of parental attitudes to continuance in child guidance treatment has also been stressed by Inman (1956) and Smigelsky (1949).

Lake and Levinger (1960) found differences in parental attitudes when they compared 50 continuers with 50 discontinuers. The parents of continuers tended to be more aware of the child's disturbance and of their own contribution to it. They were more inclined to see the problem as something for which the family as a whole was responsible and accepted that they themselves had to participate in finding a solution. They also displayed greater cooperation during interviews and tended to agree with the worker on the nature of the child's disturbance.

Levitt (1957) compared defectors with remainers on 61 variables and found 5 which seemed to differentiate between the two groups. The variables did not form a meaningful cluster, nor did there appear to be "any theoretical reason to expect them to be differentiating." In addition, the probability of finding 5 analyses out of 61 significant at the .05 level is very nearly .25. These considerations led Levitt to ascribe his results to chance. The present study attempted to anticipate this dilemma by testing a priori predictions made on the basis of logical and theoretical considerations.

METHOD

This center prepares an IBM punch-card record at the time each case is closed. The record is coded on 67 categories, ranging from objective background descriptions of the child and the family, to items dealing with developmental and health history, presenting symptoms, diagnosis, and clinic procedures.¹ The system includes every case closed since 1943 and, at the time of this investigation, contained 2,400 records. A study by Gilbert (1957) suggests that girls referred to child guidance clinics do not come from the same population as boys. Therefore, only boys' records were used for the present analysis. There were 1,497 boys among the cases. Terminators and remainers were drawn from this population.

Terminators were defined as families who had entered treatment but discontinued therapy on their own decision before five treatment interviews with the child had taken place. A total of 107 cases met these criteria.

Remainers were defined as families who had entered treatment, continued for more than 16 treatment interviews with the child, and terminated either by mutual or clinic decision. A total of 154 cases met these criteria.

Of the 67 categories of information available on each case, 27 were chosen for analysis. Those categories were selected in which the information was specific enough for research purposes and relevant to aspects of parental attitudes toward the clinic or toward the child. Specific predictions of the relations between the variables and the continuance dichotomy were made and recorded before analysis of the data. In four of the categories more than 1 prediction was made so that a total of 32 discrete predictions was put to test.

Some predictions were based on results of studies reported in the literature (Affleck & Mednick, 1959; Lake & Levinger, 1960; Levitt, 1957; Rubenstein & Levitt, 1957; Tuckman & Lavell, 1959). Other predictions were based on parents' ability to participate in treatment, which seemed relevant on the basis of clinical experience.

RESULTS

Of the 32 predictions, 9 could not be tested because insufficient information precluded a meaningful analysis. These hypotheses related to infantile behavior, tomboy or effeminate behavior, father's employment status, mother's employment status, nature of family residence, and three others related to diagnostic categories. Five of the 32 predictions were confirmed at better than the .02 level of sig-

¹ This system was instituted by William F. Finzer, Director of the center, and is maintained by Sallie Churchill. To them, and to David S. Lepson, who assisted with the analysis, the authors express their appreciation.

nificance. When 32 comparisons are possible, the probability of obtaining five differences, significant below the .02 level, by chance is less than .0001 ($CR = 5.51$). This calculation is based on the approximation described by Brožek and Tiede (1952).

For the remaining 18 predictions the results failed to reach a satisfactory level of statistical significance, although the differences were in the predicted direction in all but five of the comparisons.

Confirmed Predictions

Compared to terminators, remainers have a greater proportion of histories of developmental difficulties (complications in weaning, toilet training, delayed speech, reduced social responsiveness) ($\chi^2 = 10.12$, $p < .01$).²

The classification "Unusual Behavior" (confusion, disorientation, panic reactions, unpredictable, meaningless, and self-destructive acts) was found more frequently among the remainers than among the terminators ($\chi^2 = 11.49$, $p < .001$).

There was a higher incidence of marital disharmony (excluding divorce and separation) among the remainers than among the terminators ($\chi^2 = 5.96$, $p < .02$).

Compared to terminators, remainers contained more cases where, in addition to the child's individual therapy, both parents were seen individually for treatment. Among the terminators only one parent (usually the mother) tended to be in concurrent treatment ($\chi^2 = 15.25$, $p < .001$).

Termination was positively related to obtaining clinic service immediately after application, without having to wait for the intake interview; remainers tended to be families who had to wait for service ($\chi^2 = 29.16$, $p < .001$).

In addition to the results which confirmed specific predictions, two findings for which no hypotheses had been formulated emerged as the data were analyzed. Not having been specifically predicted, however, they are somewhat less convincing than the results cited above.

Compared to terminators, remainers have a higher incidence of specific somatic dis-

² $df = 1$ unless otherwise indicated.

orders (asthma, eczema, stuttering) as opposed to nonspecific somatic disorders (undiagnosed pains, sweating, tension) ($\chi^2 = 10.19$, $p < .01$).

When truancy was compared with other educational problems (reading and other learning difficulties, school phobia, school behavior problems) it was found more often among terminators than among remainers ($\chi^2 = 11.04$, $p < .001$).

Unconfirmed Predictions

The following relations, although not statistically significant, were in the predicted direction: The older the child, the greater the tendency to remain ($t = .25$). Families previously known to other social agencies tend to remain ($\chi^2 = .63$). Families known to juvenile court tend to terminate ($\chi^2 = 1.27$). Children with runaway problems tend to be among the terminators ($\chi^2 = 2.08$). Antisocial behavior is high among the terminators ($\chi^2 = .07$). The more advanced the mother's education, the greater the tendency to remain ($\chi^2 = 2.68$). The lower the child's intelligence, the greater the tendency to terminate ($\chi^2 = .15$). The middle income group, as opposed to the low income and high income groups, tends to remain ($\chi^2 = 3.79$).

The following comparisons were found to be in the opposite direction of predictions, but none reached a statistically significant level. It had been predicted that Negro families would tend to terminate ($\chi^2 = .99$); that fee paying cases would remain ($\chi^2 = .005$); that children with all types of somatic disorders would be among the remainers ($\chi^2 = .19$); that children with all kinds of educational problems would tend to remain ($\chi^2 = .95$); and that the more advanced the father's education the greater the tendency to remain ($\chi^2 = .04$).

DISCUSSION

Studies of abrupt termination in adult psychotherapy have shown that a patient's motivation is one factor which differentiates the terminator from the remainder (Affleck & Mednick, 1959). In child guidance therapy it is the parents' motivation which is the determining factor, but the parents must not only have the motivation to bring the child and

support his treatment experience, but they must also have the capacity to involve themselves actively in the therapy program. Premature termination tends to occur when either of these important conditions is not met.

Length of time on the waiting list between application and intake seems to be one measure of parental motivation. The parent with low motivation would not be expected to remain in clinic contact if he had to wait for any length of time. When treatment is finally started the group still on the waiting list may be assumed to contain a high proportion of well motivated individuals. Thus, there is a significantly greater proportion of waiting list cases among the remainers than among the terminators. There was a similar relation, although below the adopted level of significance, when the groups are compared on waiting from intake to diagnostic study ($\chi^2 = 1.37$, $p > .20$) and on waiting from diagnostic study to beginning of treatment ($\chi^2 = 1.79$, $p > .10$). While a waiting list is an undesirable feature of a clinic's operation, delay does, in many instances, serve as a screening device which eliminates poorly motivated patients before they get to the point of treatment.

The comparison which bears most directly on the parents' ability to involve themselves in treatment is that between families where only one parent was in concurrent treatment and those where both parents were being seen individually. When only one parent is in concurrent treatment, it is almost always the mother. The results clearly show that when the father, as well as the mother, can be involved in the treatment plan the chances that the case will terminate prematurely are greatly reduced.

It has become more universally recognized that the father plays an important and often crucial role in the treatment of children (Rubenstein & Levitt, 1957). The findings here reported lend strong support to this trend. They suggest that when the father is not in treatment, he may either actively sabotage treatment efforts or, by failing to support the mother and the child in their treatment experiences, materially undermine the process. This may be especially true in

the treatment of boys, the subjects in the present study.

Parental motivation might be expected to be a function of the distress which the child's symptoms cause the parents. Hiler (1959), studying adults in individual psychotherapy, found that patients who complain only of somatic symptoms are likely to terminate, probably because the symptom enables them to "bind" their anxiety. In child guidance therapy, the reverse seems to be true. The bizarre behavior of the mentally ill child and such specific somatic disorders as asthma may be assumed to be distressing to the parent, who would be very much aware of the disturbance and motivated to obtain help. The results of relevant comparisons strongly suggest that the more apparent the symptom, the more likely the case is to continue in treatment.

Another indirect reflection of parental involvement and motivation may be the fact that remaining was positively related to histories of developmental difficulties. A child with such a history may be assumed to have a psychological problem of long standing so that his parents might be more aware of it, more distressed by it, and thus more highly motivated to remain in treatment. In addition, a mother's report of difficulties in her child's early years may represent her awareness of her own contribution to the problem, an awareness Lake and Levinger (1960) report as a factor in continuance.

The confirmed prediction that remainers would contain proportionately more cases with marital disharmony also bears on a parent's personal involvement. The finding reflects the parents' ability to talk about a problem which concerns them. In addition, marital disharmony is a distressing personal problem for parents and one for which they might hope to find some relief through the clinic contact.

On the basis of studies reported from adult outpatient clinics (Hollingshead & Redlich, 1958; Rubenstein & Lorr, 1956), it had been predicted that certain sociocultural factors would be related to continuance in child guidance therapy. While there was a trend for the remainers to contain proportionately more Jewish than Catholic families ($\chi^2 = 2.92$, p

$= < .10$), no difference between remainers and terminators in fathers' occupational class was found ($\chi^2 = 3.72$, $df = 5$, $p = < .70$). Tuckman and Lavell (1959), who compared social status at all stages of clinic contact, also found that higher status patients were no more likely to maintain contact with the clinic than lower status patients. This would suggest that child guidance clinics are more successful than adult outpatient clinics in helping patients from the lower socioeconomic levels to remain in treatment.

It has been reported that a higher proportion of the continuers perceive the child's problem themselves, rather than through the demands made by the community, and that they desire change in themselves, as well as in their child and in their spouse (Lake & Levinger, 1960). The present results are consistent with these findings. Terminators were more often referred by an authority, such as juvenile court or school, while remainers were more often referred by friends, social agencies, or themselves ($\chi^2 = 3.77$, $p < .10$).

The incidence of truancy is significantly higher among terminators than among remainers when compared with other school problems. This result had not been specifically predicted. It seems that truancy is an expression of pathological family behavior patterns where avoidance of anxiety-arousing situations takes the form of physical departure. A family which abruptly and unilaterally terminates clinic contact and a child who truants from school may be manifesting the same basic reaction to stress.

While we were able to confirm a significant number of a priori predictions, this investigation shares the weakness of all ex post facto research on closed cases. A specific fact only becomes a research datum if the patient reports it accurately, the interviewer records it fully, and the person responsible for coding codes it correctly. It is hoped that a projected long range study will not only overcome these difficulties but also address itself to the question of treatment outcome and to the problem of parental attitudes toward the child which Lake and Levinger (1960), among others, have suspected to be an important variable in continuance of treatment contact.

SUMMARY

Families who terminated child guidance contact before the fifth treatment session were compared with families who continued treatment for a minimum of 16 interviews. Remainders had significantly more developmental difficulties, unusual behavior, marital disharmony, and specific somatic disorders. They contained significantly more cases where both parents were in concurrent treatment. Terminators had significantly more school truancy, and they had less often experienced a waiting period between application and intake. The results were discussed in terms of the importance of the parents' motivation and their ability to involve themselves in the treatment process.

REFERENCES

- AFFLECK, D. C., & MEDNICK, S. A. The use of the Rorschach test in the prediction of the abrupt terminator in individual psychotherapy. *J. consult. Psychol.*, 1959, 23, 125-128.
- BROŽEK, J., & TIEDE, K. Reliable and questionable significance in a series of statistical tests. *Psychol. Bull.*, 1952, 49, 339-341.
- GILBERT, G. M. A survey of "referral problems" in metropolitan child guidance centers. *J. clin. Psychol.*, 1957, 13, 37-42.
- HILER, E. W. Initial complaints as predictors of continuation in psychotherapy. *J. clin. Psychol.*, 1959, 15, 344-345.
- HOFSTEIN, S. Social factors in assessing treatability in child guidance. *Children*, 1957, 4, 48-53.
- HOLLINGSHEAD, A. B., & REDLICH, F. *Social class and mental illness*. New York: Wiley, 1958.
- INMAN, ANN C. Attrition in child guidance: A telephone follow-up study. *Smith Coll. Stud. soc. Wk.*, 1956, 27, 34-73.
- LAKE, MARTHA, & LEVINGER, G. Continuance beyond application interviews at a child guidance clinic. *Soc. Casewk.*, 1960, 91, 303-309.
- LEVITT, E. E. A comparison of "remainders" and "defectors" among child clinic patients. *J. consult. Psychol.*, 1957, 21, 316.
- LEVITT, E. E. A comparative judgmental study of "defection" from treatment at a child guidance clinic. *J. clin. Psychol.*, 1958, 14, 429-432.
- RUBENSTEIN, B. O., & LEVITT, M. Some observations regarding the role of fathers in child psychotherapy. *Bull. Menninger Clin.*, 1957, 21, 16-27.
- RUBENSTEIN, E. A., & LORR, M. A comparison of terminators and remainders in outpatient psychotherapy. *J. clin. Psychol.*, 1956, 12, 345-349.
- SMIGELSKY, EVA. Why parents discontinue child guidance treatment. *Smith Coll. Stud. soc. Wk.*, 1949, 19, 118-119.
- TUCKMAN, J., & LAVELL, MARTHA. Social status and clinic contact. *J. clin. Psychol.*, 1959, 15, 345-348.

(Received August 8, 1960)

CHRONICITY OF NEUROPSYCHIATRIC HOSPITALIZATION:

A PREDICTIVE SCALE

JAMES M. ANKER

Veterans Administration Hospital, Perry Point, Maryland

This paper reports an attempt to predict the length of time that newly admitted patients will stay confined in a neuropsychiatric hospital. Specifically, the primary concern is with the prediction of the relatively infrequent event of chronicity. Although most mental patients admitted are discharged as improved within a reasonable period of time, long-term neuropsychiatric chronicity has become a major problem in the field of mental health (Giedt & Schlosser, 1955; National Committee against Mental Illness, Inc., 1957). It is important to note that the probability of discharge decreases markedly with the passage of time so that, over the years, hospitals are accumulating populations which are predominantly chronic. "In the average state mental hospital, about 15% of the patients have been there less than a year; about 25% have been there between 1 and 5 years; about 60% have been there from 5 to 45 years or longer" (National Committee against Mental Illness, Inc., 1957). Despite the fact that durations of hospitalization have decreased somewhat in recent years (Harris & Norris, 1954; Kramer & Pollack, 1957) the problem of identification and special treatment of potential chronics early in the course of their hospitalization remains a most important and intriguing challenge.

Predictive studies dealing with the course of the mental patient, e.g., regarding psychotherapy, outcome of hospitalization, and likelihood of rehospitalization, generally have produced results of low predictive value (Baron, 1953a, 1953b; Bayard & Pascal, 1954; Briggs, 1958; Cole, Swensen, & Pascal, 1954; Crandall, Zubin, Mettler, & Logan, 1954; Dunham & Meltzer, 1946; Feldman, Pascal, & Swensen, 1954; Gallagher, 1954; Gildea &

Man, 1942-43; Orr, Anderson, Martin, & Philpot, 1954-55; Peterson, 1954a, 1954b; Schofield, Hathaway, Hastings, & Bell, 1954; Swensen & Pascal, 1954a, 1954b). Two studies offer more promising results. Meeker's (1958) exploratory MMPI scale to predict length of neuropsychiatric hospitalization appears to have greater value but his limited samples somewhat obscure the results. One study using demographic data has been able to predict length of neuropsychiatric hospital stay (more than or less than 90 days) with 77.2% overall accuracy (Lindemann, Fairweather, Stone, Smith, & London, 1959). A discussion of common shortcomings and difficulties in prognostic research may be found in Zubin and Windle (1954). They also point out the unusual prevalence of contradictory and low-order results.

PROBLEM

Two basic problems presented themselves for study. Improving the means of identifying potentially chronic patients among those newly admitted appeared necessary. Such identification could result in new, or at least more intensive, treatment procedures for this group and, hopefully, shorten their stay. Secondly, it seemed advantageous to develop an identification procedure which would be meaningful in relation to personality characteristics of the patient. Rather than using demographic data, for instance, this information should be more capable of suggesting specific types of treatment procedures aimed at reducing chronicity. For these reasons a personality inventory was chosen to provide the basic data for the construction of a predictive scale. The MMPI was selected because data were available. This paper reports the development of the scale.

ORIGINAL SAMPLE

Procedure. MMPI protocols at least 1 year old were taken from the records of the psychology service of a large Veterans Administration hospital. Those protocols that had been obtained over 2 months after the date of admission were not used as data. Further, because of their small number, the records of female patients were excluded. Two criterion groups were selected from these protocols; a short-term group which stayed 6 months or less and a long-term group which stayed 1 year or longer. The short-term group numbered 103 and the long-term group 63. The data from these dichotomous groups were item analyzed for every item in the MMPI item pool by chi square from a 2×2 contingency table. These computations were verified graphically by solving the chi square quadratic and generating an elliptical ABAC in the manner suggested by Andrews (1952).

Results. Fifty-five different items differentiated the criterion groups at the .05 level or less, 33 items at the .02 level or less, and 17 items at the .01 level or less. These frequencies exceeded those for a series of statistical tests at less than the .01 level (Block, 1960; Sakoda, Cohen, & Beall, 1954). The point biserial correlation between scale score (each

item arbitrarily given a possible score of 1) and the dichotomous criterion of chronicity was .53 for all 55 items. The point biserial correlation between scale score on the items differentiating at the .02 level or less and the dichotomous criterion or chronicity was .63. These values are only approximations, however, because of the discontinuity between the two samples. These original results were submitted to cross-validation procedures.

CROSS-VALIDATION SAMPLE

Procedure. MMPI protocols over 1 year old which were obtained within 2 months of admission were solicited from a number of Veterans Administration neuropsychiatric hospitals.¹ Protocols were sorted into the same criterion groups as in the original sample. The group with durations of more than 6 and less than 12 months was held for later analysis. These data were analyzed for those 55 items which differentiated the criterion groups in the original sample. The short-term group in the cross-validation sample numbered 144 and the long-term group 123. The diagnostic, racial, and age characteristics of the original and cross-validation sample combined are presented in Table 1.

Results. The cross-validating item analysis produced 21 items which differentiated between the criterion groups at approximately the .05 level or less, 11 at the .02 level or less and 9 at the .01 level or less. These items and their combined probabilities (Lindquist, 1940) are presented in Table 2.²

SCALE CONSTRUCTION

The 21 items thus selected comprised the basic scale. Data from the original and the cross-validation samples were pooled and items were weighted according to their discriminatory ability (Guilford, 1942). Only one item, number 35 in the MMPI booklet, warranted a weighted score of 2. The other 20 items in the scale were weighted 1.

Protocols from the group that stayed between 6 and 12 months were combined with

¹ The author wishes to express his gratitude to William C. Hallow, Veterans Administration Hospital, Lebanon, Pennsylvania; Burke Smith, Veterans Administration Hospital, Roanoke, Virginia; and John B. Marks, Veterans Administration Hospital, American Lake, Washington, for their cooperation in supplying the cross-validation data for this study.

² The list of the 55 items which discriminated the criterion groups in the original sample is available from the author on request.

TABLE 1
COMBINED SAMPLE CHARACTERISTICS

Diagnosis	
Schizophrenic reactions	204
Affective reactions	7
Psychoneurotic reactions	72
Paranoid state	2
Psychophysiologic reactions	4
Acute brain syndromes	8
Chronic brain syndromes	17
Personality trait disturbances	13
Personality pattern disturbances	6
Sociopathic personality disturbances	16
Transient situational personality disorders	4
Unknown	5
	358
Race	
Caucasian	309
Negro	41
Oriental	2
Unknown	6
	358
Age	
Mean	33.8
Standard deviation	9.1

Note.—The total sample of 433 was reduced to 358 through the elimination of short form and otherwise incomplete protocols.

TABLE 2
CROSS-VALIDATED SCALE ITEMS AND THEIR POOLED PROBABILITIES

Number in MMPI Booklet	Item	Pooled p Value
16	I am sure I get a raw deal from life. (T)	.0005
20	My sex life is satisfactory. (F)	.0005
35 ^a	If people had not had it in for me I would have been much more successful. (T)	.0005
42	My family does not like the work I have chosen (or the work I intend to choose for my life work). (T)	.005
54	I am liked by most people who know me. (F)	.001
60	I do not read every editorial in the newspaper every day. (F)	.0005
162	I resent having anyone take me in so cleverly that I have had to admit that it was one on me. (T)	.025
184	I commonly hear voices without knowing where they come from. (T)	.0005
252	No one cares much what happens to you. (T)	.0005
262	It does not bother me that I am not better looking. (F)	.005
265	It is safer to trust nobody. (T)	.025
278	I have often felt that strangers were looking at me critically. (T)	.01
309	I seem to make friends about as quickly as others do. (F)	.005
324	I have never been in love with anyone. (T)	.005
354	I am afraid of using a knife or anything very sharp or pointed. (T)	.005
427	I am embarrassed by dirty stories. (T)	.005
482	While in trains, busses, etc., I often talk to strangers. (F)	.005
488	I pray several times every week. (F)	.025
495	I usually "lay my cards on the table" with people that I am trying to correct or improve. (F)	.0005
540	My face has never been paralyzed. (F)	.025
556	I am very careful about my manner of dress. (F)	.005

Note.—Letters in parentheses indicate direction items are scored to reflect longer durations of hospitalization.

^a This item is assigned a scoring weight of 2. All other items have unit weight.

protocols from the item analysis criterion groups. Unfortunately, there were only six subjects in this "middle" group. The number of short form MMPIs in the combined group was insufficient to provide normative data for the short form and thus these protocols were omitted from further analysis.

The total number of protocols in this combined sample, excluding short forms, was 386. Scale scores were determined for each subject in the combined sample. In scoring protocols on the chronicity scale the number of items in the scale that were not answered was recorded. Based on the somewhat arbitrary judgment of the author, those protocols having three or more omissions were not included in the data analysis. On this basis 28 subjects were dropped leaving a total of 358, all with scores on the chronicity scale, who had various durations of stay in the hospital. The Pearson correlation between scale score and duration of stay in months was .41.

The sample was successively dichotomized, according to duration of stay, at 3 months, 6 months, 12 months, 18 months, and 23 months. No dichotomy was made at 9 months because of the uneven sampling for durations in this range. The frequency distributions of chronicity scale scores for the various duration dichotomies are presented in Figure 1. These distributions have been smoothed and adjusted for the base rate of the particular duration of stay (Cureton, 1957). These base rates were estimated by determining the duration of stay for every male patient given an MMPI after admission to a large Veterans Administration hospital between October 1954 and September 1956. Estimates were based on 240 cases.

The optimal cutting point to include all cases, after following Cureton's procedure of smoothing and adjusting the area under the curves for base rate, is the point of intersection of the two curves. These points are indi-

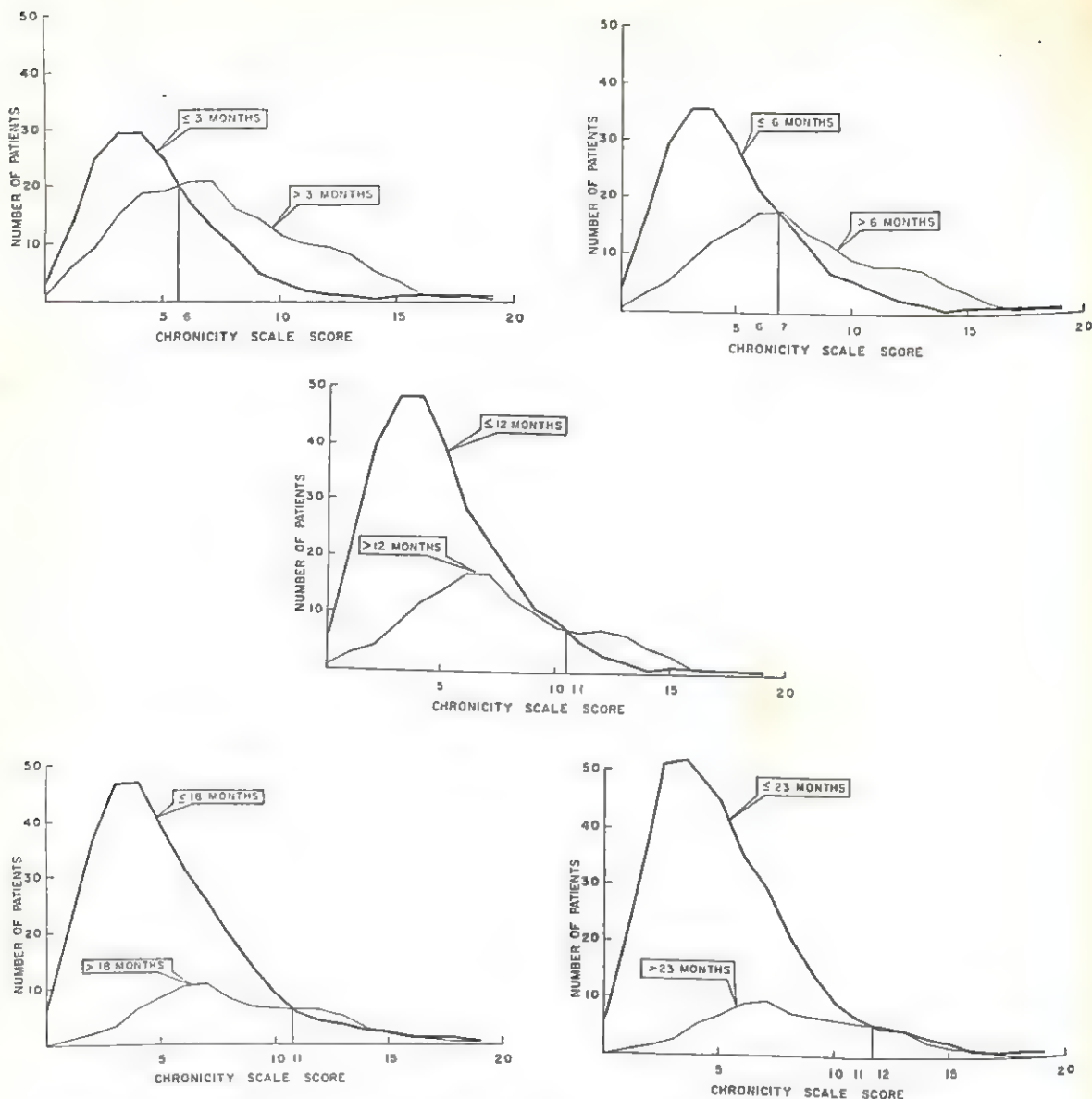


FIG. 1. Chronicity scale score distributions for dichotomized durations of hospitalization.

cated in Figure 1. Higher selectivity can be obtained, of course, by using two cutting points minimizing the area of overlap in the mid-range. This paper, because it is concerned with the total patient population, reports the findings for one cutting point and all of the sample for each dichotomy.

Use of the scale to improve decision making was the crucial consideration. This was evaluated at each of the points of dichotomy by the procedures suggested by Meehl and Rosen (1955). The predictions or "decisions" uniformly were to identify the group staying

longer rather than the shorter duration group. Table 3 presents these data and statements of the inequalities indicating whether or not use of the scale improves decision making. Improvement, or predicting better than one could with base rate information alone, occurs when the inequality is satisfied. Use of the scale aids prediction of the longer stay patient in all of the dichotomies except that for 2 years or longer. It might be mentioned that the scale improves prediction of the shorter-term patient in all cases.

Additional cross-validation data were ob-

TABLE 3
ACTUAL AND PREDICTED LENGTHS OF STAY IN MONTHS AT SEVERAL CUTTING POINTS

Predicted duration	Actual duration		Σ	$Q < \frac{P_1}{P_1 + P_2}$
	≤ 3 Months	> 3 Months		
≤ 3 Months	123.7 (70.5%)	68.0 (36.8%)	191.7	.48 < .68
> 3 Months	51.7 (29.5%)	117.0 (63.2%)	168.7	
Σ	175.4 (100%)	185.0 (100%)	360.4	
	≤ 6 Months	> 6 Months		
≤ 6 Months	169.7 (77.5%)	59.5 (42.1%)	229.2	.60 < .72
> 6 Months	49.2 (22.5%)	81.9 (57.9%)	131.1	
Σ	218.9 (100%)	141.4 (100%)	360.3	
	≤ 12 Months	> 12 Months		
≤ 12 Months	291.6 (95.5%)	105.7 (78.1%)	397.3	.69 < .83
> 12 Months	13.8 (4.5%)	29.7 (21.9%)	43.5	
Σ	305.4 (100%)	135.4 (100%)	440.8	
	≤ 18 Months	> 18 Months		
≤ 18 Months	303.0 (95.5%)	69.1 (80.7%)	372.1	.79 < .81
> 18 Months	14.2 (4.5%)	16.5 (19.3%)	30.7	
Σ	317.2 (100%)	85.6 (100%)	402.8	
	≤ 23 Months	> 23 Months		
≤ 23 Months	333.6 (95.1%)	63.7 (81.1%)	397.3	.82 < .79
> 23 Months	17.3 (4.9%)	14.8 (18.9%)	32.1	
Σ	350.9 (100%)	78.5 (100%)	429.4	

Note.— P = proportion of longer durations of stay in a specified clinical population. $Q = 1 - P$. P_1 = proportion of valid positives. P_2 = proportion of false positives. The different Σ s in each of the dichotomies is an artifact produced by proportioning the areas under the curves for base rates.

tained from the Veterans Administration Hospital, St. Cloud, Minnesota, on 204 neuropsychiatric patients having durations of hospitalization of 1 year or longer. No data for patients staying less than 1 year were available. The data obtained were evaluated against the earlier data for patients staying less than 1 year by the method described above. The cutting point determined by the earlier cross-validation data was used. Prediction of durations of greater than 1 year by the scale was better than could be expected by chance or base rate information alone. The inequality $Q < \frac{P_1}{P_1 + P_2}$, which reflects the extent decision making is improved by use of the scale, was $.69 < .75$ for this additional data, as compared to $.69 < .83$ for the original data. Thus, while the improvement in decision making (prediction) is somewhat less than the original data showed, this independent sample supports the conclusions drawn earlier. Further data on 165 neuropsychiatric admissions to the Veterans Administration Hospital, Minneapolis, Minnesota, were gathered. The median length of stay for these neuropsychiatric admissions in this general medical and surgical hospital was 46 days. The sample was dichotomized at the median, the frequency distributions of scale scores smoothed, and the intersect taken as the optimal cutting point. Predictions based on this cutting point also were better than could be

expected by chance or base rate information. The inequality $Q < \frac{P_1}{P_1 + P_2}$ in this case was $.50 < .68$.

RELATIONSHIPS WITH SIMILAR SCALES

As Zubin and Windle (1954) might predict, studies using MMPI items to predict length of hospital stay have produced remarkably little item overlap between scales purporting to do the same thing. Meeker (1958) developed a 28 item scale by item analysis of the MMPI items against a criterion of chronicity. Although his samples were relatively small and from the same area it is unusual to find that only 2 items overlapped with the 55 items discriminating in the original sample of this study. Further, both of these items were dropped from the present scale in cross-validation. A scale being developed at the Veterans Administration Hospital, American Lake, Washington, has little overlap, if any, with the present scale or with the Meeker scale.³ The lack of congruence in these independent studies, in addition to being somewhat startling, is provocative. Geographical differences might be suggested but it is not likely that this variable alone would produce such divergence. Whatever the reason, adequate sam-

³ Personal communication from John B. Marks, Veterans Administration Hospital, American Lake, Washington.

TABLE 4
 $Q < \frac{P_1}{P_1 + P_2}$ VALUES

Predicted duration of hospitalization	Meeker scale	Marks scale	Chronicity scale
> 46 Days ^a	.50 < .56	.50 < .56	.50 < .68
> 3 Months ^b	.48 < .56	.48 < .50	.48 < .68
> 6 Months ^b	.60 < .58	.60 < .61	.60 < .72
> 12 Months ^b	.69 < .76	.69 < .00 ^c	.69 < .83
> 18 Months ^b	.79 < .69	.79 < 1.00 ^c	.79 < .81
> 23 Months ^b	— ^d	.82 < .88 ^e	.82 < .79
> 12 Months ^e			.69 < .75

^a Sample of 165 neuropsychiatric admissions to the Minneapolis Veterans Administration General Medical and Surgical Hospital.
^b Data from original pooled cross-validation sample, $N = 259$.
^c Inequality not interpretable because cell frequencies approach zero.
^d No cutting point obtained because frequency distributions did not intersect.
^e Data from 204 neuropsychiatric admissions to the Veterans Administration Hospital, St. Cloud, Minnesota, that stayed 1 year or longer.

pling and cross-validation appear to be the immediate course of action. In the present study an attempt was made to obtain adequate sample sizes and cross-validation. Regarding the sample, however, the number of subjects who had durations of hospitalization between 6 and 12 months is decidedly deficient. Combined sample size was adequate but undoubtedly could be improved.

These other scales were evaluated as to their capacity to improve decision making in comparison with the present scale on the same data samples whenever the other scale scores could be obtained. Table 4 presents the $Q < \frac{P_1}{P_1 + P_2}$ values for each of the scales on a number of predictions and samples. The present scale uniformly improves decision making more than the other scales, does so more consistently, and for predictions of longer durations of hospitalization.

One might suspect a relationship between a scale predicting response to psychotherapy and one predicting length of hospitalization. There is only one item, however, which overlaps between Barron's Ego Strength scale and the chronicity scale. For this one item the direction of scoring which would contribute to a favorable prognosis for psychotherapy from Barron's scale would, in the chronicity scale, contribute to the prediction of long stay. It seems obvious that the two scales are not measuring the same underlying variable(s).

DISCUSSION

The results of this study have provided a 21 item scale which improves prediction of the longer stay patient. It can discriminate meaningfully for durations of hospitalization up to 18 months. Although accuracy of prediction decreases as duration increases, it does appear that the scale may have practical value, particularly for isolating the potentially chronic group with a minimum of "false positives." Such predictions undoubtedly could be improved by combining probabilities with a separate predictor such as the demographic instrument developed by Lindemann et al. (1959). It is also possible prediction might have been more accurate had slightly less rigorous standards been used in the selection of items. Although the .05 level was used in

the item analysis, the pooled probabilities were considerably lower. Predictive efficiency may have been enhanced had the .05 level, for instance, been used for the pooled probabilities, thus generating a longer scale.

The substance of this study is the scale that has been developed empirically. While it will serve to improve identification of the chronic patient, at this point it provides only interesting clues about self-reported personality characteristics underlying chronicity. The scale will be factor analyzed to define the "roots" of chronicity insofar as they are reflected in these MMPI items. These results will be presented in a subsequent paper. It was the intent of this study to produce a predictive scale which might be independent of diagnosis, at least to the extent that it would eventually permit isolation of personality factors influencing chronicity. Although it is most probable that these factors themselves will not be entirely independent of diagnosis it appeared fruitful to investigate them separately.

Although considerable care was taken in the construction and cross-validation of the present scale, in view of the current contradictory findings in this area it must be viewed as tentative and properly subject to further evaluation and modification. It should prove to be of some immediate interest to the clinician addressing himself to the study of neuropsychiatric chronicity if it is used with the proper reservations. Eventually it may increase our understanding of chronicity and consequently improve our ability to study and impede it.

SUMMARY

The MMPI item pool was item analyzed against a dichotomous criterion of neuropsychiatric hospital chronicity. The 55 items which were found to discriminate the criterion groups in the original sample were cross-validated on the pooled data from three separate Veterans Administration neuropsychiatric hospitals. A 21 item scale was generated which was able to predict the "long-stay" patient at various dichotomies in duration of stay better than one could by chance or by base rate information. Although it may be of some immediate value, because of the prevalence

of contradictory findings in this area, the scale should be used with caution until further verification and modification. The scale will be factor analyzed to provide some notion of the underlying "roots" of chronicity, at least insofar as they can be tapped by self-report on a questionnaire. These results will be presented in a later publication.

REFERENCES

- ANDREWS, T. G. Mass data formulation of the 2×2 contingency table for chi-square. Unpublished manuscript, University of Maryland, 1952.
- BARRON, F. An ego-strength scale which predicts response to psychotherapy. *J. consult. Psychol.*, 1953, 17, 327-333. (a)
- BARRON, F. Some test correlates of response to psychotherapy. *J. consult. Psychol.*, 1953, 17, 235-241. (b)
- BAYARD, JEAN, & PASCAL, G. R. Studies of prognostic criteria in the case records of hospitalized mental patients: Affective expression. *J. consult. Psychol.*, 1954, 18, 122-126.
- BLOCK, J. On the number of significant findings to be expected by chance. *Psychometrika*, 1960, 25, 369-380.
- BRIGGS, P. F. Prediction of rehospitalization using the MMPI. *J. clin. Psychol.*, 1958, 14, 83-84.
- COLE, MARY E., SWENSEN, C. H., & PASCAL, G. R. Prognostic significance of precipitating stress in mental illness. *J. consult. Psychol.*, 1954, 18, 171-175.
- CRANDALL, A., ZUBIN, J., METTLER, F. A., & LOGAN, NOREEN. The prognostic value of "mobility" during the first two years of hospitalization for mental disorder. *Psychiat. Quart.*, 1954, 28, 185-210.
- CURETON, E. E. Recipe for a cookbook. *Psychol. Bull.*, 1957, 54, 494-497.
- DUNHAM, H. W., & MELTZER, B. N. Predicting length of hospitalization of mental patients. *Amer. J. Sociol.*, 1946, 50, 123-131.
- FELDMAN, DOROTHY A., PASCAL, G. R., & SWENSEN, C. H. Direction of aggression as a prognostic variable in mental illness. *J. consult. Psychol.*, 1954, 18, 167-170.
- GALLAGHER, J. J. Test indicators for therapy prognosis. *J. consult. Psychol.*, 1954, 18, 409-413.
- GIEDT, F. H., & SCHLOSSER, J. R. Movement of patients through a neuropsychiatric hospital. Unpublished manuscript, Medical Library, Veterans Administration Hospital, Perry Point, Maryland, 1955.
- GILDEA, E. F., & MAN, EVELYN B. Methods of estimating capacity for recovery in patients with manic-depressive and schizophrenic psychoses. *Amer. J. Psychiat.*, 1942-43, 99, 496-506.
- GUILFORD, J. P. *Fundamental statistics in psychology and education*. New York: McGraw-Hill, 1942.
- HARRIS, A., & NORRIS, VERA. Changes in duration of stay of mental hospital patients suffering from functional psychoses during the past 20 years. *J. ment. Sci.*, 1954, 100, 241-249.
- KRAMER, M., & POLLACK, E. S. Problems in the interpretation of trends in the population movement of the public mental hospitals. Paper read at American Public Health Association, Cleveland, November 15, 1957.
- LINDEMANN, J. E., FAIRWEATHER, G. W., STONE, G. B., SMITH, R. S., & LONDON, I. T. The use of demographic characteristics in predicting length of neuropsychiatric hospital stay. *J. consult. Psychol.*, 1959, 23, 85-89.
- LINDQUIST, E. F. *Statistical analysis in educational research*. New York: Houghton Mifflin, 1940.
- MEEHL, P. E., & ROSEN, A. Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychol. Bull.*, 1955, 52, 194-216.
- MEEKER, F. O. An exploratory study of specific and general personality dimensions related to length of hospitalization among psychiatric patients. Unpublished master's thesis, University of California, 1958.
- NATIONAL COMMITTEE AGAINST MENTAL ILLNESS, INC. *What are the facts about mental illness?* Washington, D. C.: NCAMI, 1957.
- ORR, W. F., ANDERSON, RUTH B., MARTIN, MARGARET P., & PHILPOT, D. F. Factors influencing discharge of female patients from a state mental hospital. *Amer. J. Psychiat.*, 1954-55, 111, 576-582.
- PETERSON, D. R. The diagnosis of subclinical schizophrenia. *J. consult. Psychol.*, 1954, 18, 198-200. (a)
- PETERSON, D. R. Predicting hospitalization of psychiatric out-patients. *J. abnorm. soc. Psychol.*, 1954, 49, 260-265. (b)
- SAKODA, J. M., COHEN, B. H., & BEALL, G. Test of significance for a series of statistical tests. *Psychol. Bull.*, 1954, 51, 172-175.
- SCHOFIELD, W., HATHAWAY, S. R., HASTINGS, D. W., & BELL, DOROTHY M. Prognostic factors in schizophrenia. *J. consult. Psychol.*, 1954, 18, 155-166.
- SWENSEN, C. H., JR., & PASCAL, G. R. Duration of illness as a prognostic indicator in mental illness. *J. consult. Psychol.*, 1954, 18, 363-365. (a)
- SWENSEN, C. H., JR., & PASCAL, G. R. Prognostic significance of type of onset of mental illness. *J. consult. Psychol.*, 1954, 18, 127-130. (b)
- ZUBIN, J., & WINDLE, C. Psychological prognosis of outcome in the mental disorders. *J. abnorm. soc. Psychol.*, 1954, 49, 272-281.

(Received August 11, 1960)

TRAIT JUDGMENT OF PHOTOGRAPHS AND ADJUSTMENT OF COLLEGE STUDENTS

JAY L. CHAMBERS

Charles L. Mix Memorial Fund, Inc., Americus, Georgia

A variety of judgment tasks have discriminated those with severe mental or emotional disorders from less severely disturbed personalities (Chambers, 1956, 1957; Cooper, 1960). Judgment tasks applied to clinical populations have varied from psychophysical judgments of length of lines, density of dots, etc., to projective types of judgment (e.g., judging personality traits from photographs, classifying ambiguous stimuli such as inkblots, etc.). Sarbin and Hardyck (1955) defined the latter type of judgment as "conformance" and applied the concept to tasks where there is no criterion of judgment validity other than agreement with a specified norm group.

In the present study a conformance type of judgment task was given to college students. It was hoped that the task might be sufficiently sensitive to discriminate between well adjusted and poorly adjusted students, even though poorly adjusted college students would not generally compare in severity of pathology with the psychotic groups previously discriminated by judgment tasks.

METHOD

Picture Identification Test (PIT). At the beginning of a school year, all students at Georgia Southwestern Junior College were given the group form of the PIT (Chambers & Broussard, 1960). There were 256 male and 186 female students tested. For part of this test the subjects were given 10 cards on each of which were eight head and shoulder photographs of individuals of the same sex as the subject. With each card the subject was given a list of 21 brief behavior descriptions representing 21 needs of the Murray (1953) need system. The subject was instructed to select a picture to match each description using any picture as often as he desired. The subject thus made 10 selections for each of the 21 needs, making a total of 210 choices.

Scoring. Norms for the popularity of judgment choices were based on the choices of 100 randomly selected students of each sex from the general student population. The score for any selection was determined by the number of those in the norm group making that selection. Twenty-one judgment subscores were obtained by summing the 10 choice scores for each need. A total judgment score was obtained by summing all 210 choice scores. The scores of the subjects of the norm groups were adjusted to eliminate their own contribution to the norm figures. The 21 raw judgment subscores and the total judgment score were converted to standardized scores derived from the total student population and based on an eight-point scale with the mean separating points four and five on the scale.

Subjects. At the end of the school year in which the PIT had been administered, each member of the faculty was asked to select the 10 best adjusted students of each sex and the 10 most poorly adjusted students of each sex from those of his acquaintance. The faculty members were instructed to judge students on the basis of emotional stability and maturity rather than on intellectual ability. None of the faculty knew of the PIT results. Fourteen faculty members felt they had enough general contact with the student body to make selections. Those students selected by three or more of the faculty judges as best adjusted were chosen for a "well adjusted" group and those receiving three or more choices as most poorly adjusted were selected for a "poorly adjusted" group. This procedure provided 15 boys and 16 girls for the poorly adjusted group and 16 boys and 17 girls for the well adjusted group.

RESULTS

Judgment subscores of well adjusted and poorly adjusted boys were compared for differences by means of the *t* test. The well adjusted group showed a higher mean degree of judgment conformity on all 21 subscores and they differed significantly (at the .05 level or better) from the poorly adjusted group on 19 of the 21 subscores.

A similar analysis was applied to the results of the well adjusted and poorly adjusted

TABLE 1

TOTAL JUDGMENT SCORES OF WELL ADJUSTED AND POORLY ADJUSTED STUDENTS

Total Judgment Score	Well Adjusted		Poorly Adjusted	
	Boys	Girls	Boys	Girls
1	4	4	1	
2	3	4		
3	2	1	1	2
4	5	5	2	3
5	1		4	4
6	1	1		3
7		1	3	1
8		1	4	3

* Dotted line indicates mean for general student population.

female groups. The well adjusted girls showed a higher mean degree of judgment conformity on all but two subscores (n Dominance and n Infavoidance) and they were significantly more conforming in judgment than the poorly adjusted group on 11 of the 21 subscores.

Table 1 presents the total judgment scores for all four groups. A test of the difference between the combined male and female well adjusted groups ($N = 33$) and the combined male and female poorly adjusted groups ($N = 31$) yielded a t of 11.09, significant at the .001 level. A cutting point at the mean for the general population of students ($\bar{X} = 4.5$) classified 71% of the poorly adjusted students below the mean and 85% of the well adjusted students above the mean.

Further comparisons revealed that the well adjusted groups scored significantly higher than the poorly adjusted groups on the Princeton Scholastic Aptitude Test ($t = 2.64$, $p < .02$). This finding raised the possibility that the PIT judgment score might be correlated with the SAT. A Pearson r of .28 was found between the PIT total judgment score and the SAT. Based on an N of 212, this correlation was significant at the .01 level but was not high enough to indicate use of the PIT judgment score as an index of scholastic aptitude or to suggest that prediction of general adjustment to a college environment based on the PIT judgment score would yield highly similar results to prediction based on the SAT.

DISCUSSION

Evidence from the present study and from those previously cited points to impaired or deviant perceptual judgment among the emotionally or mentally disturbed. In an attempt to account for the relationship between adjustment and judgment, Sarbin and Hardyck (1955) reasoned that a lack of perceptual conformance would increase the probability of a deviant behavioral response since behavioral reactions are dependent on perceptual reactions. Mental and emotional disorders are defined by deviant behavior; therefore there is a direct connection between poor perceptual conformance and maladjustment, according to their theory.

The above theory does not state what conditions underlie poor perceptual conformance. The conditions could, of course, be due to brain damage or be of genetic origin.

There is a further possibility that judgment may be distorted by emotions and attitudes. Jurors and judges are dismissed from cases where they hold certain attitudes affecting the case. Perceptual judgments may be subverted to maintain what Morgan (1956) has termed the self-preservation of attitudes and beliefs. According to this theory, there is a selective process whereby evidence favorable to existing attitudes is registered while negative evidence is excluded. It is obvious that one who persistently modifies perception rather than belief, when the two are incompatible, will eventually find himself at odds with others and with reality. It would seem better, from a mental health viewpoint, that perception should be uncensored and that attitudes should be influenced by perceptual evidence.

Instances of poor perceptual conformance may provide indicators of specific psychological conflicts. An inspection of the data of the present study showed that, for a given individual, perceptual conformance varied considerably from one need to the next. Future research might profitably attempt to determine whether specific problem or conflict areas can be diagnosed by differences in perceptual conformance scores of the various needs as measured by the PIT.

SUMMARY

The group form of the Picture Identification Test was administered to all students at a small state junior college at the beginning of a school year. Each subject received 21 judgment subscores and a total judgment score based on his matchings of photographs of people with Murray need descriptions. At the end of the school year, 33 students were selected by faculty members to comprise an emotionally well adjusted student group and 31 students were chosen as showing poor emotional adjustment.

The boys chosen as well adjusted made significantly higher scores (more popular matchings) than the poorly adjusted boys on 19 of the 21 judgment subscores. The well adjusted girls made significantly higher scores than did the poorly adjusted girls on 11 of the 21 judgment subscores.

On the total judgment score, the difference between the combined well adjusted and combined poorly adjusted groups was highly reliable ($t = 11.09$, $p < .001$). Using the general population mean as a cutting point, 78% of the subjects could be classified by the total judgment score of the PIT in agreement with faculty selections of well adjusted and poorly

adjusted students. A low but significant r of .28 ($p < .01$; $N = 212$) was found between the PIT total judgment score and the Princeton Scholastic Aptitude Test.

The results were discussed in relation to Sarbin's theory of perceptual conformance and Morgan's theory of perceptual selectivity as a means of insuring the self-preservation of existing attitudes and beliefs.

REFERENCES

- CHAMBERS, J. L. Perceptual judgment and associative learning ability of schizophrenics and nonpsychotics. *J. consult. Psychol.*, 1956, 20, 211-214.
- CHAMBERS, J. L. Trait judgment of photographs by neuropsychiatric patients. *J. clin. Psychol.*, 1957, 13, 393-396.
- CHAMBERS, J. L., & BROUSSARD, L. J. Need attitudes of normal and paranoid schizophrenic males. *J. clin. Psychol.*, 1960, 16, 233-237.
- COOPER, RUTH. Objective measures of perception in schizophrenics and normals. *J. consult. Psychol.*, 1960, 24, 209-214.
- MORGAN, C. T. *Introduction to psychology*. New York: McGraw-Hill, 1956.
- MURRAY, H. A. *Explorations in personality*. New York: Oxford Univer. Press, 1953.
- SARBIN, T. R., & HARDYCK, C. C. Conformance in role perception as a personality variable. *J. consult. Psychol.*, 1955, 19, 109-111.

(Received August 14, 1960)

THREE RORSCHACH SCORES INDICATIVE OF SCHIZOPHRENIA

IRVING B. WEINER

University of Rochester School of Medicine and Dentistry

The use of Rorschach summary scores in differential diagnosis, though widespread in clinical practice, has seldom been justified by the research literature. Attempts to cross-validate checklists such as the Miale and Harrower-Erickson (1940) signs of neurosis and the Davidson Rorschach adjustment scale (Davidson, 1950) have typically been unsuccessful (Berkowitz & Levine, 1953; Corsini & Uehling, 1954). Rieman (1953) combed the literature for Rorschach signs supposedly discriminatory between neurotics and schizophrenics. Of 86 such "indicators" he evaluated, only 5, or barely a chance expectation, discriminated significantly between the neurotic and schizophrenic subjects he studied.

Results such as these have dampened research interest in Rorschach summary scores as diagnostic indicators, if the paucity of recent literature in this area may be taken as a criterion. However, related work has pointed up potential sources of error in the determination of diagnostic signs which often have not been taken into account. From papers by Cronbach (1949), Fiske and Baughman (1953), and Knopf (1956) several guidelines for the empirical derivation of Rorschach signs may be abstracted: (a) the number of independent statistical tests applied should be kept at a minimum; (b) nonparametric methods of statistical inference should be used; (c) the response total should be controlled; and (d) Rorschach scores and dependent variables must be kept independent. The present study consists of the empirical selection of three Rorschach signs associated with severity of psychopathology in an exploratory study and two subsequent attempts to cross-validate these summary scores as indicators of schizophrenia.

EXPLORATORY STUDY

The exploratory sample consisted of all adult patients in a 6-month period from the psychiatric services of a general hospital for whom scorable Rorschach protocols were available, with the exception of those patients whose primary diagnosis was organic rather than functional in nature. The subjects' hospital and clinic records were utilized to assign them to one of three diagnostic categories: *neurosis*, which included cases of conversion hysteria, obsessive-compulsive neurosis, anxiety reaction, and neurotic depressive reaction; *character disorder*, which applied to instances of personality trait and personality pattern disturbances; and *psychosis*, which included schizophrenic, psychotic depressive, and involutional psychotic reactions. The diagnostic criterion was the label assigned to the patient during the period of psychiatric evaluation in which the Rorschach had been given. For hospital patients, these labels were the discharge diagnoses; for clinic patients, the recorded consensus of an intake committee was used.

The Rorschach protocols for this sample, which numbered 71, had previously been scored for the presence of a number of signs under investigation in a related context. Three of the signs unexpectedly appeared to differentiate among the diagnostic categories. To the extent that the labels neurosis, character disorder, and psychosis represent a continuum of increasing disturbance, severity of illness was significantly associated with tendencies to (a) give 1 or 2 *CF*, (b) have a Sum *C* between 1.5 and 3.0, and (c) give at least 1 *CF* or *C* response with no *C'* responses. Table 1 indicates the number of subjects in each diag-

TABLE 1

FREQUENCIES OF EXPLORATORY SUBJECTS OF DIFFERENT DIAGNOSES WITH A GIVEN NUMBER OF SCHIZOPHRENIC INDICATORS

Diagnostic Category	Number of Indicators*			
	0	1	2	3
Neurosis	18	2	4	2
Character disorder	8	4	5	7
Psychosis	5	2	7	7

* $\chi^2 = 13.581$ with 6 *df*; $p < .05$.

nostic category who received from none to all three of these signs. Subsequent analysis revealed no significant differences between the three diagnostic groups in mean age or in mean or variance of response total. Furthermore, each group was composed of approximately two-thirds females and one-third males, thus mitigating any influence of sex difference.

CROSS-VALIDATION STUDIES

Procedure

Two separate samples were utilized to evaluate the concurrent validity of a checklist comprised of the above three summary scores. These samples, A and B, were again selected from case files and consisted, respectively, of 52 patients from the 6 months following and 89 patients from the 12 months preceding the exploratory period. Since the psychotic group in the exploratory study had contained few nonschizophrenic patients, only schizophrenics were included in the psychotic group for the cross-validating samples. Sample A contained 16 neurotics, 18 character disorders, and 18 schizophrenics; for Sample B the totals were 27, 31, and 31. The total population ranged in age from 15 to 62 with a mean of 29.62 years and had given a mean number of 23.23 responses to the Rorschach. Within each sample there were no significant differences among the three diagnostic categories in mean and range of age or in mean and variance of response total. Each group had approximately 50% males, the six groups ranging from 44 to 61% males.

It was impossible to determine the extent to which the Rorschach findings had influenced the diagnostic label attached to the patients, thus contaminating the data. However, it was the investigator's hypothesis that even in cases where the decision to designate a patient schizophrenic had depended entirely on the Rorschach findings, it was unlikely that the three signs under consideration here—1 or 2 *CF*, Sum *C* 1.5 to 3.0, and *C* or *CF* without *C'*—had contributed prominently to the psychologist's opinion that the

patient's Rorschach was consistent with the presence of schizophrenia. To test this ancillary hypothesis, five staff psychologists, who had in fact done or supervised the testing of more than three-fourths of the subjects in the study, were asked to select from the following list of 10 Rorschach variables 6 which they felt were of either primary or secondary importance in the assessment of schizophrenia: (a) number and/or % *Dd*; (b) Sum *C*; (c) number confabulated and contaminated *Rs*; (d) number *CF*; (e) number and/or % *P*; (f) *F* + %; (g) number and/or quality *M*; (h) number *C*, *Csymb*, and *Cnam*; (i) amount bizarre content; and (j) number *C'*.

Results

Table 2 indicates the number of subjects with different diagnoses in Samples A and B who received a given number of the signs. The degree of association between diagnostic category and number of signs, as estimated by χ^2 , is significant beyond the .001 level of confidence for both samples. Further investigation of Table 2 reveals that the major difference exists between the schizophrenic group on one hand and the neurotics and character disorders on the other. It may be seen that in Sample A, 87% of the neurotics and 78% of the character disorders received one sign or less, while only 22% of the schizophrenics had less than two signs; for Sample B the corresponding percentages are 81, 68, and 22. Chi square values computed for these grouped data were significant beyond the .001 level for both samples. Comparisons of the relative frequency with which the neurotics and character disorders received one or less or two or

TABLE 2

FREQUENCIES OF SUBJECTS OF DIFFERENT DIAGNOSES IN SAMPLES A AND B WITH A GIVEN NUMBER OF SCHIZOPHRENIC INDICATORS

Diagnostic Category	Number of Indicators**							
	Sample A				Sample B			
	0	1	2	3	0	1	2	3
Neurosis	12	2	0	2	16	6	4	1
Character disorder	12	2	4	0	11	10	9	1
Schizophrenia	3	1	5	9	3	4	15	9

** For Sample A $\chi^2 = 28.881$ with 6 *df*; $p < .001$. For Sample B $\chi^2 = 30.072$ with 6 *df*; $p < .001$.

more signs yielded χ^2 values smaller than 1.00 in each sample.

The data were analyzed further to determine if the three signs were contributing differentially to the above results. Six 2×3 contingency tables were used to assess the association between the three diagnostic categories and the presence or absence of each sign for Samples A and B. The obtained χ^2 values were all significant beyond the .01 level.

The ratings by the psychologists of which Rorschach variables are prominent in their evaluation of possible schizophrenia also yielded a clear-cut result. Six of the elements were endorsed as being of primary or secondary importance by at least four of the five raters, and a seventh was chosen by three raters. The remaining three variables were not selected by any of the psychologists as being of either primary or secondary importance in the diagnosis of schizophrenia. These three were Sum *C*, number *CF*, and number *C'*, the variables under investigation in this study.

The fact that positive results were thus obtained for indicators not endorsed by the testers might suggest either that the test reports carried little weight in the final psychiatric diagnosis or that the psychologists' diagnostic conclusions were little influenced by their expressed views on indicators of schizophrenia. To deal with this question, an additional analysis was undertaken with Sample A. It was found that the psychologists' test reports and the finally established diagnoses concurred in the presence or absence of schizophrenia in 90% of the cases, which casts doubt on the first of the above alternatives. Investigation of those of the endorsed signs which can be objectively scored revealed that the schizophrenic group had a significantly ($p < .05$) lower *F* + % and tended ($p < .10$) more frequently to have given 1 or more pure *C* response and/or 1 or more *M*- response and/or less than 1 *M* response than the non-schizophrenic groups. However, only five of the schizophrenics in Sample A had *F* + %s below 70, only five failed to give *M*, and only four gave any *M*- or *C*; only four had as many as two of these signs. Therefore, these scores, though differentiating between groups, were not present with sufficient frequency

to have much facilitated the psychologist's evaluation in the individual case. Furthermore, no relationship could be discovered between diagnostic category and *P*, *P*%, or *Dd*%. Consequently, it would seem in many cases that the psychologists, sometimes of necessity and sometimes of choice, had based their diagnostic impressions on indicators other than those they endorsed in the checklist.

DISCUSSION

The data clearly indicate that likelihood of being diagnosed schizophrenic in the psychiatric setting studied is associated with tendencies to give 1 or 2 *CF*, a Sum *C* from 1.5 to 3.0, and *C* or *CF* without *C'* on the Rorschach. The fact that these three schizophrenic indicators were suggested by an exploratory study and cross-validated at highly significant levels of confidence in two subsequent studies makes it unlikely that they merely reflect chance fluctuations. The built-in controls for age, response total, and sex eliminate several frequent sources of error. The author's emphasis on nonexistent contaminating variables is felt necessary because the results cannot readily be explained in terms of prevalent Rorschach theory. There is no literature which suggests that these three signs are associated with schizophrenia, and the rating sheet used in this study demonstrates that psychologists, at least those in the setting where this research was done, do not consciously utilize these signs in evaluating schizophrenic potential. That positive results were achieved seems, in view of the congruence between the examiners' judgments and the diagnostic labels of the subjects, attributable to the finding that some of the traditional signs of schizophrenia they endorsed were often ignored by them and others occurred with insufficient frequency to figure prominently in their individual diagnostic conclusions. The nature of the design does not allow further inference concerning the validity of these traditional signs, however.

A tentative rationale for the findings may be offered. The three signs taken together represent some minimal use of chromatic color without use of achromatic color. Individuals

who avoid color almost entirely or who use both chromatic and achromatic color freely would not receive the signs. It may follow that, within a patient population, those persons who use color freely are displaying the emotional lability frequently associated with repressive defense, while those who avoid color are manifesting the isolation of affect which accompanies intellectual defenses. The remaining patients, having neither pattern, might be viewed as lacking a stable defensive structure and being prone to schizophrenic reactions under stress. Other explanations are certainly possible, but confirmation of any hypotheses suggested by the data would require additional research.

It is the author's feeling that even with present knowledge the data have value for the clinician. It should be noted that the signs were not derived and cross-validated by comparing blatant schizophrenics with normals. Psychological test data gathered in such studies are often useful for making discriminations only in situations where the discrimination is so obvious that psychological tests are superfluous. The subjects in this study were selected from actual case files and typically had been referred because the diagnosis was not clear. Hence they constitute the type of diagnostic problem with which the clinician must deal in his daily practice, and it is within such a population that the three signs were so extremely in accord with the diagnoses finally established. The signs by no means identified all the schizophrenic patients. However, the data seem to justify their inclusion among other Rorschach variables commonly construed as relating to the presence of a schizophrenic disorder.

SUMMARY

Three Rorschach signs—1 or 2 *CF*, Sum *C* between 1.5 and 3.0, and *C* or *CF* without *C'*—were found significantly associated with severity of psychopathology in an exploratory study. In two cross-validating studies, with 52 and 89 subjects, each of these signs was received significantly more frequently by schizophrenic patients than by neurotics and character disorders. The design contained controls for age and sex and for mean and variance of Rorschach response total. A tentative rationale for the results is offered, and it is felt that the data recommend these signs for inclusion among Rorschach criteria for the presence of schizophrenia.

REFERENCES

- BERKOWITZ, M., & LEVINE, J. Rorschach scoring categories as diagnostic "signs." *J. consult. Psychol.*, 1953, 17, 110-112.
- CORSINI, R. J., & UEHLING, H. F. A cross-validation of Davidson's Rorschach adjustment scale. *J. consult. Psychol.*, 1954, 18, 277.
- CRONBACH, L. J. Statistical methods applied to Rorschach scores. *Psychol. Bull.*, 1949, 46, 393-429.
- DAVIDSON, HELEN H. A measure of adjustment obtained from the Rorschach protocol. *J. proj. Tech.*, 1950, 14, 31-38.
- FISKE, D. W., & BAUGHMAN, E. E. Relationships between Rorschach scoring categories and the total number of responses. *J. abnorm. soc. Psychol.*, 1953, 48, 25-32.
- KNOFF, I. J. Rorschach summary scores in differential diagnosis. *J. consult. Psychol.*, 1956, 20, 99-104.
- MIALE, FLORENCE R., & HARROWER-ERICKSON, MOLLIE R. Personality structure in the psychoneuroses. *Rorschach res. Exch.*, 1940, 4, 71-74.
- RIEMAN, G. W. The effectiveness of Rorschach elements in the discrimination between neurotics and ambulatory schizophrenics. *J. consult. Psychol.*, 1953, 17, 25-31.

(Received August 24, 1960)

ANOTHER LOOK AT MMPI PROFILE TYPES IN MULTIPLE SCLEROSIS

HAROLD GILBERSTADT AND EDWIN FARKAS¹

Veterans Administration Hospital, Minneapolis, Minnesota

Canter (1951) reported a descriptive study of MMPI profiles of patients with multiple sclerosis (MS). Although recognizing the limitations of the approach, he calculated the average MMPI profile for this group of 33 World War II veteran patients and inferred from the mean profile for the group that the typical personality configuration in MS included a reaction to the stress of the illness with depression and its accessory symptoms.

When depression is a major variable under study, the averaging of MMPI profiles can obscure important profile differences especially if subsamples are combined, one of which has very high *D* scores and the other of which has very low *D* scores. This is a particularly important consideration in the study of an illness such as MS in which both reactions of high depression and low depression because of denial and repression have been observed. If discrete profile types were to be produced reflecting each of these reaction types, such an important difference would be cancelled out and masked by averaging.

The aim of the present study was to attempt to check Canter's MMPI findings and to answer the following questions: (a) Can a typical response to MS in the direction of depression be inferred from the MMPI? (b) Among patients with neurological lesions, are MMPI profiles indicative of depression more common in MS than in other conditions? To investigate the latter question, a control group of neurological patients was selected who had suffered brain injury from external causes.

A second purpose of this paper is to report on significant relationships between MMPI

profile characteristics and illness and demographic variables which became apparent when profiles were studied as depressed and non-depressed types rather than as average group profiles.

METHOD

Samples

The MS sample consisted of 25 male veterans hospitalized on the Neurology Service of the Minneapolis Veterans Administration Hospital who had received a medical diagnosis of MS. The control group consisted of 25 male veterans from Neurology with a medical diagnosis of traumatic brain injury. Except for three cases in the MS group, the samples represented all of the cases with these diagnoses who had been administered the MMPI and Wechsler-Bellevue on the Neurology Service during the years 1949 to 1952. Three MS cases currently on the Neurology Service were tested and included in the analysis to increase the size of the sample.

Age, IQ, and duration of illness data are reported in Table 1. The group differences in mean age and mean IQ were not statistically significant. The mean duration of illness of the Minneapolis MS sample was 58 months compared to 30 months for the control group; this difference is reliable ($p < .001$). Canter's MS group would appear to be comparable to the present MS group in age and duration of illness although there may be differences between the groups in socioeconomic status. Also, Canter studied outpatient veterans while the present sample of veterans was hospitalized.

Table 1 also reports the MMPI average *T* scores of the three samples. All three of the mean profiles were highly similar in shape; Canter's mean profile shows higher elevations on *Hs*, *D*, and *Hy* than the other two profiles. In addition, the Minneapolis MS sample is significantly higher on *D* and *Sc* than the Minneapolis control group. The magnitudes of the latter differences are small.

Procedure

The Minneapolis MS and control samples were subdivided on the basis of MMPI scores. The MMPI profile groupings were based on the following considerations. First, since there is a relatively high in-

¹ We wish to thank William Schofield, Jan Duker, and Irving Gottesman for suggestions about the preparation of this manuscript.

TABLE 1

MEAN AGE, IQ, DURATION OF ILLNESS, AND MMPI T SCORES FOR MINNEAPOLIS
MULTIPLE SCLEROSIS AND CONTROL SAMPLES AND CANTER MULTIPLE SCLEROSIS SAMPLE

Sample	Age	IQ	Duration (months)	L	F	K	Il5	D	Hy	Pd	Mf	Pa	Pt	Sc	Ma	Si
Minneapolis MS sample (<i>N</i> = 25)	32.9	102.4	58	55.4	55.0	59.4	71.6	69.6	68.0	60.0	51.5	54.7	62.3	65.0	54.6	52.0
Minneapolis control sample (<i>N</i> = 25))	30.6	102.9	30.2	54.2	53.5	56.2	68.7	62.3	65.6	56.4	51.4	52.3	59.9	60.1	57.1	49.2
Canter MS sample (<i>N</i> = 33)	32	—	48	55	55	58	81	79	75	59	55	53	63	64	55	—

cidence of one scale being elevated over *T* score 70 even in normal samples (Hathaway & Meehl, 1951) it was decided to classify the profiles as abnormal if two or more scales were elevated beyond the normal limits. Second, since the investigation focused on depression, it seemed rational to cut the samples on the basis of elevation on the *D* scale. Third, inspection of the set of profiles suggested that the abnormal-normal dichotomy and the depressed-nondepressed categories would describe much of the relevant information obtained in a majority of the profiles.

The following categories and rules were adopted:

1. Normal Depressed: Not more than one scale over 70.

$$D > 60$$

2. Normal Nondepressed: Not more than one scale over 70.

$$D < 60$$

3. Abnormal Depressed: Two or more scales over 70.

$$D > 80$$

4. Abnormal Nondepressed: Two or more scales over 70.

$$D < 80$$

RESULTS

Table 2 shows the resulting distribution of the profile types in the Minneapolis MS and

TABLE 2

FREQUENCY OF PROFILE TYPES IN MULTIPLE SCLEROSIS
AND CONTROL GROUPS

Profile Type	Multiple Sclerosis	Control
Normal Depressed	5	5
Normal Nondepressed	7	8
Abnormal Depressed	8	1
Abnormal Nondepressed	5	11

control samples. The chi square test revealed the distributions of the two samples to be reliably different ($\chi^2 = 7.76$, $p < .05$).

Table 3 shows the mean age, IQ, duration of illness, and MMPI *T* scores for each of the four profile types in the MS and control groups. Within the four MS profile types, significant *F* ratios were obtained for age ($p < .05$), IQ ($p < .01$), and duration of illness ($p < .01$). From the Sheffé test (Sheffé, 1953), which was applied to make all possible comparisons of means, it appears that (a) the Normal Nondepressed type was significantly younger than the Abnormal Depressed type, (b) the Normal Nondepressed type had a significantly higher mean IQ than the Abnormal Depressed type, (c) the Normal Nondepressed type had a significantly shorter duration of illness than any of the three other types, and (d) the Abnormal Nondepressed type also had a significantly higher mean IQ than the Abnormal Depressed type.

No significant *F* ratios were found in the analysis of the control group.

There appeared to be an interaction between the age, IQ, and duration of illness variables within the profile types in the MS sample. Compared to the mean score for the total Minneapolis MS sample, every case of the Normal Nondepressed profile type had lower than average age and higher than average IQ and shorter than average duration of illness except for one case where age was at the mean. Seven of the eight cases of the Abnormal Depressed type had IQ scores below the mean of the total MS group.

TABLE 3
MEAN AGE, IQ, DURATION OF ILLNESS, AND MMPI T SCORES FOR PROFILE SUBTYPES OF MULTIPLE SCLEROSIS AND CONTROL SAMPLES

Profile Type	Age	IQ	Duration (months)	L	F	K	Il _s	D	Il _y	Pd	Mf	Pa	Pl	Sc	Ma	Si
Multiple sclerosis sample																
Normal Depressed (N = 5)	32.2	97.8	74.4	54.6	52.0	59.8	65.0	64.0	61.6	54.0	44.4	40.4	57.8	58.6	53.2	53.0
Normal Nondepressed (N = 7)	26.9	112.1	14.3	53.0	51.1	63.7	60.1	52.1	61.0	38.9	53.1	54.1	53.6	56.4	55.7	43.3
Abnormal Depressed (N = 8)	38.9	92.4	73.5	58.9	62.9	54.0	81.6	91.6	73.9	66.3	50.8	64.1	74.5	76.5	53.5	60.4
Abnormal Nondepressed (N = 5)	32.6	109.2	77.8	53.8	50.6	61.6	78.0	64.6	74.8	57.8	57.4	54.6	59.4	64.8	56.0	49.6
Control sample																
Normal Depressed (N = 5)	38.8	105.2	34.0	55.6	53.8	57.2	71.4	69.6	66.2	56.4	42.4	52.4	60.8	59.8	60.0	51.8
Normal Nondepressed (N = 8)	30.9	103.9	30.0	54.0	52.5	52.5	55.5	49.3	54.8	53.3	50.5	51.4	47.5	52.3	58.8	45.4
Abnormal Depressed (N = 1)	31	114	48	50	50	44	88	87	78	50	65	56	81	74	63	72
Abnormal Nondepressed (N = 11)	26.5	100.2	27.1	54.0	54.4	59.5	75.4	66.2	72.5	59.4	54.9	52.5	66.5	64.7	54.1	48.8

DISCUSSION

From the present findings, it would seem safe to state that there is a reaction to MS in the direction of severe depression more frequently than in a neurological disorder such as traumatic brain injury. Thirty-two percent of the MS sample obtained MMPI profiles of the Abnormal Depressed type compared to 4% of the control sample. The longer duration of illness in the MS group could be a factor contributing to the greater incidence of such profiles in the MS sample. Vieg (1947) reported that the most important mechanism used by patients in adjusting to MS was repression, but that as the disease progressed this kind of equilibrium could not be maintained and the patients gave up their front of being emotionally well adjusted and happy. It would not seem warranted to call dysphoria the typical reaction to MS since only a third of the sample fell into the Abnormal Depressed category. Rather, it would appear that there are several kinds of reaction to the disease. Philippopoulos, Wittkower, and Cousineau (1958) found a broad range of personality structures in a group of 40 MS patients studied through interview and psychological tests.

Most studies of MS (Grinker, Hamm, & Robbins, 1948; Langworthy, 1948; Philippopoulos et al., 1958; Sugar & Nadell, 1948; Vieg, 1947) seem to show a strong thread of consistency in reporting that repression, inability to express hostile feelings directly, overconformity, and/or hysteroid traits in general are common in patients with MS either after, or both before and after the outward manifestations of the MS symptoms. It should be noted in passing that similar personality traits have been reported in illnesses which are generally considered to be psychosomatic in nature such as the recent investigation of ulcer patients by Marshall (1960). Some support for the observation that repression and denial are preferred mechanisms of at least some MS patients could be found in the present MMPI data but the same evidence could be found for the control sample. It would be impossible to distinguish cause from effect in any case.

On the one hand, it would seem that it is unwarranted to make sweeping generalizations about typical emotional reactions to MS, but on the other hand, it is not necessary to adopt a narrow idiographic point of view. There is evidence that multivariate classification on demographic variables such as age, intelligence, duration of illness, and perhaps sex, is needed to establish most effectively the relationships between illness characteristics and personality variables. In the present study, it would appear that severe depression is more likely to occur in less intelligent male MS patients and that younger, more intelligent male patients with shorter duration of illness are more likely to have a repertory of responses which enable them to avoid depression. These relationships are not evident in the control sample. The simplest explanation for the absence of such relationships in the control sample would be in terms of the disabling and progressive nature of MS which might evoke responses which would not be called forth by a brain injury that was more specific in its consequences and not progressive. However, this explanation would not seem to be completely satisfactory and the possibility that at least some of the MS patients would not overlap the control patients on significant personality characteristics cannot be ruled out.

There is evidence in the present data of the need for longitudinal analysis in the study of MS. Data concerning preillness educational and occupational attainment were incomplete but suggested the need for further study of the possibility that the Abnormal Depressed MS cases had lower IQs and the Normal Non-depressed MS cases had higher IQs which antedated the onset of diagnosable MS symptoms. The changes in patients over time after the development of symptoms is illustrated by the case of a 31-year-old college graduate with an IQ of 129 who received a diagnosis of MS shortly after the development of symptoms of the disease. At that time he obtained an MMPI profile with a peak score on *Hy* at 65 which would place him in the Normal Non-depressed profile category. Eight years later he obtained an IQ of 104 and an MMPI profile of the Abnormal Depressed type with a peak on *D* of 109.

In general, it might be concluded that it is dangerous to use any single measure of psychological status, presuming that such status is sensitive to a condition of illness, when it is either established or likely that the psychological variable is correlated with demographic and other nonillness variables which have not been controlled.

SUMMARY

MMPI and Wechsler-Bellevue performance of 25 patients with MS and 25 control patients with traumatic brain injuries indicated:

1. There was a reaction of severe depression in 34% of the MS group compared to 4% of the control group as inferred from MMPI profile types. This supported Canter's previous findings in part but did not appear to warrant the generalization that the typical response to MS is in the direction of depression.

2. MS patients with profile types indicating severe depression had lower IQs than average for the group of MS patients.

3. Younger, more intelligent MS patients with shorter duration of illness than average for the group tended to obtain nondepressed profiles.

4. The MS group differed from the control group in that no relationship of profile type with IQ, duration of illness, or age was found in the control group. This was considered most likely to be due to the differing nature of the diseases but alternative explanations could not be ruled out.

5. The importance of a longitudinal approach to MS was indicated.

6. The need was pointed out for control of all possible nonillness variables in order to be able to conclude that any single psychological characteristic is sensitive to a condition of illness.

REFERENCES

- CANTER, A. H. MMPI profiles in multiple sclerosis. *J. consult. Psychol.*, 1951, 15, 253-256.
- GRINKER, R., HAMM, G. C., & ROBBINS, F. P. Some psychodynamic factors in multiple sclerosis. *Proceedings of the twenty-eighth Annual Meeting of the Association for Research in Nervous Mental Diseases*, 1948.
- HATHAWAY, S. R., & MEEHL, P. E. The Minnesota Multiphasic Personality Inventory. In, *Military clinical psychology*. (Department of the Army Technical Manual TM8:242, Department of the Air Force Manual AFM 160-145) Washington, D. C.: United States Government Printing Office, 1951.
- LANGWORTHY, D. R. A survey of the maladjustment problems in multiple sclerosis and the possibilities of psychotherapy. *Proceedings of the twenty-eighth Annual Meeting of the Association for Research in Nervous Mental Diseases*, 1948.
- MARSHALL, SIMONE. Personality correlates of peptic ulcer patients. *J. consult. Psychol.*, 1960, 24, 218-223.
- PHILIPPOPOULOS, G. S., WITTKOWER, E. D., & COUSINEAU, A. The etiologic significance of emotional factors in onset and exacerbations of multiple sclerosis. *Psychosom. Med.*, 1958, 20, 458-474.
- SCHEFFÉ, H. A method for judging all contrasts in analysis of variance. *Biometrika*, 1953, 40, 87.
- SUGAR, C., & NADELL, R. Mental symptoms in multiple sclerosis. *J. nerv. ment. Dis.*, 1948, 98, 267-280.
- VIEG, M. J. Clinical investigation into the psychological aspects of multiple sclerosis. Unpublished doctoral dissertation, University of Minnesota, 1947.

(Received August 25, 1960)

DICHOTOMOUS EVALUATIONS IN SUICIDAL INDIVIDUALS

CHARLES NEURINGER

Suicide Prevention Center, Los Angeles, California

It need not be pointed out that suicide is a leading cause of death in the United States and that it constitutes a serious mental health problem. Research is slow and difficult because of a number of complex methodological difficulties, among which is the problem of an adequate definition of the suicidal rubric. So many different kinds of behaviors and experiences come under this heading that it is difficult to perceive a common core of psychological characteristics that can be denoted as "suicidal" and easily discriminated from other pathological states.

Edwin S. Shneidman (1960) has proposed such a core of characteristics of thinking in suicidal individuals. He posited that the neurotic-suicidal individual is the perpetrator of thought distortions, the presence of which leads to a high probability of suicide occurring. One of these proposed types of thought distortions was the tendency to think in terms of absolute value dichotomies.

By Dichotomous Evaluative Thinking is meant the polarization of thought into an extreme double bind value system (e.g., "good vs. bad, right vs. wrong, beautiful vs. ugly"). The polarization is considered to be extreme in that the object of thought is considered as "all or mostly good" or "all or mostly bad," with very little modulation of the object into finer discriminations, commonly called "shades of grey."

Shneidman feels that rigid adherence to extreme Dichotomous Evaluative Thinking can lead to situations that are lethal (e.g., if an individual is dissatisfied with his life and does not find it wholly acceptable to him, he does not think of alternate ways of changing his life but of death as the only alternative).

Dichotomous thinking seems to be an "either-or" kind of value thinking and not an "and" kind of thinking. Shneidman con-

cluded that the extreme dichotomous thinker is trapped in a double bind and must always embrace one of the extremes. The implication is that if one adheres to a strict dichotomy of thought, then one has few or no degrees of freedom with which to maneuver. Alternatives cannot be perceived and the situation becomes unresolvable, thus leading to ideas of escape through death.

Should Dichotomous Evaluative Thinking be a distinguishing characteristic of suicidal thinking, this would lend support to the view that there are common characteristics that are organized into a core personality, which could be called the "suicidal personality."

PROBLEM

The general question that the following research attempted to answer was whether suicidal individuals can be differentiated from (a) another emotionally disturbed group and (b) from a normal control group, as far as the presence and extent of extreme Dichotomous Evaluative Thinking was concerned.

METHOD

The Semantic Differential method, which was developed by Charles Osgood (1957), was selected as a means of studying the hypothesis that Dichotomous Evaluative Thinking is a characteristic of suicidal thought.

Dichotomous thinking of an extreme kind should reflect itself in extreme evaluations of the concept being tested. If suicidal subjects evaluate more dichotomously than the control subjects, they should score a concept as being "more good, more beautiful, more unpleasant and more sad" than the control group evaluations of the same concept.

Another reflection of dichotomous thinking would be found in the amount of difference in values between two concepts that are semantically opposite such as *love* and *hate* or *God* and the *devil*. Greater Dichotomous Evaluative Thinking of an extreme kind in the suicidal individuals should reflect itself in the perception of greater value differences on the

same pair of opposing concepts when compared with the control groups.

Consistent with Osgood's suggestion concerning the use of the Semantic Differential when studying values, only those scales high on the evaluation factor were utilized. In this study nine scales (good-bad, dirty-clean, nice-awful, unpleasant-pleasant, fair-unfair, worthless-valuable, happy-sad, dishonest-honest, and beautiful-ugly), and 18 concepts (democracy, death, God, honor, communism, mother, success, love, life, murder, devil, father, myself, shame, failure, other people, hate, and suicide) were used. The concepts that were chosen reflect important people in one's life, political systems, emotional states, behavioral acts, theological entities, etc. These kinds of concepts were selected on the assumption that they would tend to elicit strong evaluative reactions in most people. It was felt that more impersonal concepts such as steel, apple, or skyscraper would not evoke measurable evaluative thinking. It is difficult to categorize steel as being good or bad, honest or dishonest, happy or sad, fair or unfair, etc. The concepts that were chosen had less veridical objectivity and therefore should be open to greater value interpretation.

Twelve of the concepts were organized into six semantically opposed pairs (God-devil, life-death, honor-shame, success-failure, love-hate, and democracy-communism) in order to test for the amount of perceived value difference between the members of the paired concepts.

Scoring

The scoring of a subject's Semantic Differential for value extremeness was accomplished by assigning a score of 3 for extreme judgments like "very good" or "very bad," a score of 2 for moderate judgments like "moderately beautiful" or "moderately ugly," a score of 1 for judgments such as "mildly worthless" or "mildly valuable." A score of 0 was assigned to the midpoint rating. The more extreme the judgment, the higher the score on the scale. The mean value extremeness score per scale was recorded for the subject. Each subject made 162 scale judgments.

The scoring of a subject's Semantic Differential for value differences was accomplished by comparing the rating differences between a pair of oppositional concepts scale by scale. The difference score per scale was the number of rating spaces separating the two judgments, including the scored spaces on the same scale. Complete opposition such as *life* being "very good" and *death* as being "very bad" received a score of 7. Complete identity received a score of 1. The greater the amount of disparity in scale ratings, the greater the value difference score. The mean value difference score per scale was recorded for the subject. There were 54 scale difference scores available for each subject. The greater the value difference score, the greater the value disparity between the opposing concepts.

Congruent with the hypothesis that suicidal individuals evaluate more dichotomously than other people,

it would be expected that the suicidal subjects in this study would have greater value extremeness and difference scores than the control subjects.

Subjects

The subjects were gathered from five Veterans Administration hospitals and one large metropolitan general hospital. Four of the hospitals were in the Los Angeles area and two were in the Missouri-Kansas area. Four of the hospitals were general medical and surgical and two were neuropsychiatric.

Three groups of 15 subjects each were utilized in the present study. The first of these groups was composed of individuals who had made a serious attempt at killing themselves (S group). The second group were subjects suffering from marked psychosomatic difficulties (PS group) and the last group was composed of normal hospitalized patients (N group).

All the subjects were native-born, Caucasian males between the ages of 21 and 55. They were of normal intelligence as defined by the Information subtest of the Wechsler-Bellevue Intelligence Scale, Form I (Wechsler, 1944). (The S, PS, and N groups earned mean Information subtest scores of 16.4, 17.2, and 17.6, respectively. The corresponding standard deviations were 4.1, 2.9, and 3.4. An analysis of variance was carried out and an *F* ratio of 1.46 was found which for 2 and 42 degrees of freedom was not significant at the .05 level of confidence.) None of the subjects were psychotic and they were all in good enough mental and physical shape to partake in the research. All the subjects were hospitalized at the time of the Semantic Differential administrations. The suicidal subjects were hospitalized because of their suicide attempts, and the psychosomatic subjects were in the hospital receiving treatment for their physical difficulties. Hospitalized normal patients were used as a control for the effects of hospitalization.

None of the subjects received any kind of psychiatric or psychological treatment previous to partaking in the research project. Great care was taken to select subjects who had had no electroshock, ataratic drugs, individual or group psychotherapy.

Besides these general characteristics of the subjects, the suicidal subjects were chosen on the basis of their having made a bona fide suicidal attempt. The establishment of the suicide attempt was made by (a) the verbal admission of the patient and (b) the presence of some objective evidence of self-destruction such as a high barbiturate level in the blood, noxious chemical substances lavaged from the stomach, and deep surgical wounds on the body. In addition, the suicidal subjects carried a diagnosis of neurosis as defined by the American Psychiatric Association diagnostic manual (1952).

The psychosomatic subjects were categorized by (a) no history of suicidal threat or attempt and (b) a diagnosis of psychosomatic disability as defined by the above cited manual.

The normal subjects were defined by (a) no evidence of suicidal threat or attempt and (b) no evidence of emotional disturbance. The absence of emo-

tional difficulties was approximated with critical items from the Cornell selective indices (Mittlemann & Brodman, 1946). The normal patients were in the hospital only for medical and surgical problems of a transient nature, such as appendectomies, tonsil removals, and minor fractures. Patients that suffered from physical difficulties such as skin diseases or internal gastric disorders were not utilized because of the possibility of unknown psychosomatic involvement. Patients who had remained in the hospital longer than 1 month were eliminated since there might be unknown psychological factors which were propedeutic to their staying in the hospital longer than prescribed by the nature of their physical disabilities.

RESULTS

The results of the study suggest that there is little difference between the suicidal and psychosomatic subjects as far as extreme Dichotomous Evaluative Thinking is concerned. However, a significant difference was found between the two emotionally disturbed groups and the normal hospitalized subjects.

Extremeness Scores

The percentage of the different kinds of value extremeness responses made by the three groups of subjects is presented in Table 1. For the high value extremeness category (Number 3, indicating extreme judgments such as "very good, very bad," etc.), the suicidal group earned the highest percentage of responses (71%), while the normal subjects earned the lowest percentage of responses (52%). This relationship is reversed for the low value extremeness category (Number 0, indicating neutral or equal judgments such as "equally good and bad, equally honest and dishonest," etc.). For these modulated ratings, the normal group earned the highest per-

TABLE 1

PERCENTAGE OF RESPONSES MADE ON EACH VALUE EXTREMENESS CATEGORY BY THE SUICIDAL, PSYCHOSOMATIC, AND NORMAL SUBJECTS ON THE SEMANTIC DIFFERENTIAL TEXT

Group	Extremeness Category				Total %
	3	2	1	0	
S group	71	15	4	10	100
PS group	67	8	10	15	100
N group	52	15	8	25	100

TABLE 2

MEANS AND STANDARD DEVIATIONS OF THE VALUE EXTREMENESS SCORES PER SCALE FOR THE SUICIDAL, PSYCHOSOMATIC, AND NORMAL SUBJECTS

Measure	S Group	PS Group	N Group
Mean	2.45	2.26	1.93
SD	.37	.39	.46

centage (25%), and the lowest percentage (10%) was obtained by the suicidal subjects. The distribution of extremeness scores for the psychosomatic group lies closer to the suicidal group than to the normal group. The means and standard deviations of the extremeness score per scale are presented in Table 2.

It appears that all the subjects used extreme value judgments and their judgments were distributed in roughly the same manner. But the suicidal and psychosomatic subjects made more extreme value judgments than the normal subjects. It was hypothesized that the suicidal subjects, if they used Dichotomous Evaluation Thinking to a greater degree than the control subjects, would assign more extreme value ratings to the Semantic Differential scales than the control subjects. A simple analysis of variance of the scale extremeness scores was computed and the summary is presented in Table 3. An F ratio of 6.62 was found, which was significant at the .01 level of confidence. The source of the differences between the means was traced by the Tukey method (1949) and it was found that both the suicidal and psychosomatic groups earned significantly higher value extremeness scores than the normal group. There was no statistical difference between the two emotionally disturbed groups.

TABLE 3

ANALYSIS OF VARIANCE OF THE VALUE EXTREMENESS SCORE PER SCALE FOR THE SUICIDAL, PSYCHOSOMATIC, AND NORMAL GROUPS

Source	df	Mean Square	F
Between	2	1.06	6.62*
Within	42	.16	
Total	44		

* Significant at the .01 level of confidence.

TABLE 4

PERCENTAGE OF VALUE RESPONSES MADE ON EACH DIFFERENCE CATEGORY BY THE SUICIDAL, PSYCHOSOMATIC, AND NORMAL SUBJECTS ON THE SEMANTIC DIFFERENTIAL TEST

Group	Difference Category							Total %
	7	6	5	4	3	2	1	
S group	50	16	6	13	6	4	5	100
PS group	39	6	14	22	6	5	8	100
N group	26	10	8	26	7	6	17	100

Difference Scores

The percentage of scale value difference scores earned by the three groups of subjects is presented in Table 4. The left hand part of the table represents the more extreme value difference categories, while the right hand part of the table represents the less extreme value difference categories. The suicidal subjects placed 50% of their responses in the most extreme value difference category, while only 26% of the normal subjects' responses were found there. The least extreme value difference category finds the normal group contributing 17% of their responses, while the suicidal group contributed 5% of their responses. The psychosomatic group distribution lies in between the suicidal and normal groups and is closer to the former. The means and standard deviations of the difference score per scale are presented in Table 5.

From the tabular material, it can be seen that all the subjects made a great number of value differentiations between the opposing concepts. The only difference seems to lie in the extent of the differences. It was hypothesized that the suicidal subjects, if they used Dichotomous Evaluative Thinking to a greater

TABLE 5

MEANS AND STANDARD DEVIATIONS OF THE VALUE DIFFERENCE SCORES PER SCALE FOR THE SUICIDAL, PSYCHOSOMATIC, AND NORMAL SUBJECTS

Measure	S Group	PS Group	N Group
Mean	5.59	5.03	4.34
SD	.80	.80	1.00

degree than the control subjects, would perceive greater value differences between the paired oppositional concepts than the control subjects. A simple analysis of variance was carried out on the scale difference scores and a summary of the analysis is presented in Table 6. An F ratio of 6.77, significant at the .01 level of confidence, was found. The source of the significant difference was traced by the Tukey method (1949) and it was found that the suicidal and psychosomatic groups earned significantly greater value difference scores than the normal group. There was no statistical difference between the two emotionally disturbed groups.

TABLE 6

ANALYSIS OF VARIANCE OF THE VALUE DIFFERENCE SCORE PER SCALE FOR THE SUICIDAL, PSYCHOSOMATIC, AND NORMAL GROUPS

Source	df	Mean Square	F
Between	2	5.15	6.77*
Within	42	.76	
Total	44		

* Significant at the .01 level of confidence.

DISCUSSION

The evidence gathered in this study did not support the contention that Dichotomous Evaluative Thinking was primarily or solely a characteristic of the thinking of the suicidal individuals utilized in this study. The suicidal subjects tended to score their Semantic Differentials in such a manner as to reflect the presence of dichotomous evaluations. However, both control groups showed this tendency, and the psychosomatic subjects showed just as many dichotomous evaluations as the suicidal subjects. None of the three groups of subjects demonstrated exclusive functioning as far as Dichotomous Evaluative Thinking was concerned. The differentiation among the three groups seemed to be the amount of Dichotomous Evaluative Thinking that the subjects used.

It should be stressed that the findings of this study should not be generalized to other aspects of thinking, but should be restricted

to the cognitive organizations of values. It is not known whether emotionally disturbed individuals utilize more dichotomous evaluations than normal people when dealing with impersonal objects such as "tin cans or soup spoons." Neither is it known whether Dichotomous Evaluative Thinking extends into areas of cognitive organization where values do not play a major role (e.g., syllogistic reasoning, mathematics, etc.).

Even though it has been suggested by Osgood (1957) that most everything is thought of in terms of some kind of value orientation, it cannot be said that Dichotomous Evaluative Thinking would have been found if scales high on the Activity and Potency factors were used in the present study. It may well be that steel is stronger or that carbonated water is "fizzier" for emotionally disturbed individuals in comparison to normal people, but data that would substantiate such a contention are not available.

The concept of Dichotomous Evaluative Thinking should be restricted to objects of thought which are pregnant with personal meanings and therefore elicit strong value reactions in most people.

If, on the basis that self-destruction is a drastic solution to personal difficulties, it is assumed that the suicidal individuals were under the greatest amount of stress, and that the normal subjects had comparatively the lowest level of stress, then the results of this study might be compatible with the interpretation that stress causes Dichotomous Evaluative Thinking to take on pathological and nonadaptive characteristics. A cognitive mode of thinking that is normal under normal conditions becomes a pathological caricature when the person is under great psychic stress. It would appear that excessive utilization of Dichotomous Evaluative Thinking is a common characteristic of neurosis and emotional disturbance. It is felt that suicide is not caused by Dichotomous Evaluative Thinking, but rather that its pathological and excessive utilization accompanies personality decompensation, one other manifestation of which might be suicide.

SUMMARY

The hypothesis that suicidal individuals think more dichotomously than other emotionally disturbed people and normal subjects was tested by the use of the Semantic Differential test.

Two measures of Dichotomous Evaluative Thinking were devised. They were (a) a value extremeness score (extremeness of value judgments on the Semantic Differential scales) and (b) a value difference score (the amount of value difference perceived between two opposing concepts per scale).

Three groups of 15 subjects each were utilized. They were a group of individuals that attempted suicide, psychosomatic patients, and normal hospitalized patients. All the subjects were Caucasian native-born males. Each subject was administered a Semantic Differential form consisting of 18 concepts with nine scales for each concept.

The results indicate that for the value extremeness and difference scores, the suicidal and psychosomatic subjects earned higher scores than the normal subjects. No statistical differences were found between the suicidal and psychosomatic groups on the two types of scores.

It was concluded that Dichotomous Evaluative Thinking seems to be a common characteristic of emotionally disturbed persons and was not an exclusive factor in the thinking of suicidal individuals.

REFERENCES

- AMERICAN PSYCHIATRIC ASSOCIATION, Committee on Nomenclature and Statistics. *Mental disorders*. Washington, D. C.: APA, 1952.
- MITTLEMANN, BELA, & BRODMAN, K. The Cornell indices and Cornell Word Form: Construction and standardization. *Ann. NY Acad. Sci.*, 1946, **46**, 575.
- OSGOOD, C. *The measurement of meaning*. Urbana, Ill.: Univer. Illinois Press, 1957.
- SHNEIDMAN, E. S. Psycho-logic: A personality approach to patterns of thinking. In J. Kagen & G. Lesser (Eds.), *Contemporary issues in apperceptive fantasy*. Springfield, Ill.: Charles C Thomas, 1960.
- TUKEY, J. W. Comparing individual means in the analysis of variance. *Biometrics*, 1949, **5**, 99-114.
- WECHSLER, D. *The measurement of adult intelligence*. (3rd ed.) Baltimore: Williams & Wilkins, 1944.

(Received August 26, 1960)

EVALUATION OF PSYCHOTHERAPY AS AN ADJUNCT TO INSULIN-COMA THERAPY¹

PHILIP ROOS

Board for Texas State Hospitals and Special Schools

There are three main viewpoints prevalent today concerning the value of psychotherapy as an adjunct to the somatic therapies: (a) some form of psychotherapy must be combined with somatic therapy if genuine improvement of the patient is to occur (Palmer & Riepenhoff, 1950); (b) it is undesirable to combine psychotherapy with somatic therapies since somatic therapies may actually render the patient less accessible to psychotherapy (Frank, 1950; Rosen, 1953; Sullivan, 1940); and (c) psychotherapy contributes little or nothing to the treatment of the patient over and above such methods as insulin-coma therapy (Kalinowski & Hoch, 1952).

The present study sought at least a partial answer to the following question: to what extent does psychotherapy contribute to the improvement of schizophrenic patients who are undergoing insulin-coma treatment? Studying the effects of psychotherapy in conjunction with insulin therapy and the total treatment program that is associated with it furnished an unusually rigid test of the value of psychotherapy. In this case, the comparison was not between a group receiving psychotherapy and a group receiving no specific treatment, but it was between a group receiving insulin therapy as a part of a general treatment program and a group receiving this plus psychotherapy.

METHOD

Essentially, the research design consisted of comparing the improvement made by two groups of

¹ This paper is based upon a doctoral dissertation done at the University of Texas, January 1955. The writer wishes to express his appreciation to Wayne Holtzman for his guidance and enthusiastic support. Recognition is also due the staff of the Waco Veterans Administration Hospital where this study was conducted.

schizophrenics undergoing deep insulin-coma therapy. The experimental group differed from the control group in only one variable; namely, the addition of individual and group psychotherapy to the total insulin treatment program. The progress made by the patients was assessed by diverse techniques judged to be objective, quantifiable, and meaningful. The judgments and ratings of improvement were conducted in such a way as to eliminate contamination of possible bias on the part of the investigators. The psychotherapy used with the experimental subjects was studied by means of recordings, therapists' notes, and quantified therapists' ratings.

Subjects

Over a 9-month period complete data were gathered on 19 experimental and 18 control cases. All cases were male veterans in the Veterans Administration Hospital, Waco, Texas. As patients were admitted to the insulin service, they were randomly assigned to the experimental and control groups. No significant differences were found between the two groups in age, education, or chronicity of neuropsychiatric condition.² Pretreatment ratings on scales based on psychiatric interviews, observed ward behavior, and psychological tests clearly demonstrated the presence of psychopathology in all patients prior to therapy.³

Insulin Treatment Service

The insulin service accommodated a maximum of 16 patients so that each group consisted of no more than 8 patients at any given time. All the research patients lived on the same ward and were exposed to the same ward personnel, psychiatrists, social worker, and psychologists, as well as to the same general environment. They were housed in a special wing which provided treatment room, day room, and sleeping quarters.

Typically, the patients underwent insulin coma 5 days a week. In rare cases insulin coma was combined with electric shock treatment, or the comas

² A detailed description of these subjects is available in Roos (1955).

³ The scales and rating procedures are described in a later section of this paper.

were interspersed with shock treatments.⁴ These modifications in treatment were determined by the psychiatrist who kept a very close watch on each patient. The overall treatment time was set at 10 weeks by the psychiatrist so that all patients would have the same amount of somatic therapy and the same length of time on the intensive treatment service.

Occupational therapy, recreational therapy, and social service functions were available to all patients. For the duration of the research project, however, the social service worker limited himself to discussing superficial reality problems with the insulin patients in order to avoid possible contaminating effects.

Evaluative Criteria

Three types of evaluative measures were used: (a) scales based on a psychiatric interview and observed behavior on the ward, (b) scales based on psychological test data, and (c) length of stay in the hospital following termination of the insulin coma treatment.

Multidimensional Scale for Rating Psychiatric Patients (MSRPP). The MSRPP is divided into two sections, one referring to a psychiatric interview and the other focusing upon ward behavior as judged by nurses and psychiatric aides (Lorr, 1953). Preliminary studies utilizing this scale report satisfactory reliability and validity (Lorr, 1953; Lorr, Jenkins, & Holsopple, 1954). The scale yields an overall morbidity score indicative of degree of pathological variation from "normal" behavior.

All those using the scales were carefully instructed in the proper rating procedure. Common rating errors, such as halo effect, central tendency and position effect, were discussed in group meetings. The psychiatrist in charge of the insulin service was requested to use the MSRPP Interview scales to record his impressions of the patient based upon an interview prior to onset of insulin treatment and again ten weeks later when treatment was terminated. The psychiatrist was in close contact with each patient throughout his treatment as well as for several weeks prior to its initiation. He was unaware, however, of which patients received psychotherapy and was careful not to elicit material from them which might reveal this information to him, thus preventing the possibility of unconscious bias influencing his judgments.

The MSRPP Ward Observation scales were presented to the nurses, aides, occupational therapists, and recreational workers, and they were asked to rate each patient biweekly. These workers had close daily contact with the patients so that these ratings should be revealing of the patients' behavior in social situations and on the ward. These raters were given the impression that the research was concerned only with intensive study of patients undergoing insulin treatment. Thus, independent behavioral ratings

obtained from the ward personnel provided unbiased criteria of therapeutic success.

Psychological Tests. Each patient was administered a battery of psychological tests before the beginning of treatment and again upon its completion. These tests were given by a psychologist other than the experimenter, who was unaware of whether a given patient was in the control or experimental group. The battery included the following group-administered tests: (a) the Grayson Perceptualization test, (b) the Kent EGY intelligence test, (c) the Human Figure Drawing Test, (d) the Bender-Gestalt, (e) a sentence completion test, and (f) a modification of the Wechsler-Bellevue Picture Completion Test. All research patients were also given the Rorschach test before and after treatment. This test was administered individually with a standard inquiry following the technique outlined by Klopfer and Kelley (1946). Each protocol was then scored, the psychogram was plotted, and the main ratios computed.

Test interpretation scales intended to quantify significant aspects of personality which might be expected to show changes as a result of psychotherapy were devised by the investigator. Two sets of rating scales were developed. One set, consisting of five steps, was designed to assess the patient's functioning in certain areas at the time of testing. The other set, consisting of seven steps, was aimed at estimating degree of change taking place during a lapse of time in the same areas of functioning. The first set, referred to as Cross-Sectional Rating scales, was applicable to a single test. The second set, referred to as Comparison Rating scales, was applicable to combined pre-posttests. The names of the seven scales follow: ⁵ Intellectual Efficiency, Emotional Functioning, Psychopathological Characteristics, Interpersonal Efficiency (Nature of Social Contacts), Interpersonal Efficiency (Need Gratification), Self-Attitudes, and Psychosexual Level.

Three clinical psychologists served as raters; all had considerable experience in psychodiagnostics. The scales were discussed in detail with them. Only after the scales were clearly understood and considered unambiguous by all raters were they applied to the research cases.

Cross-Sectional Ratings. All identifying data were removed from the test items as well as any information pertaining to whether the given material was obtained before or after treatment. The group tests were treated as one unit, and the Rorschach was treated as a separate unit. The raters were first presented with a set of group tests from 10 patients, or a set of 10 Rorschachs, consisting of pre- and posttests and both control and experimental cases. They then rated each Rorschach and each set of group tests (which were treated independently and without knowledge of which group tests belonged with which Rorschach) on the Cross-Sectional Rating scales. Following this procedure, which was repeated with several sets of test protocols for 10 patients at a

⁴ The treatment data on each patient are summarized in Roos (1955).

⁵ The actual operational definitions for each step on the scales can be found in Roos (1955).

time, each rater was handed a set of 10 Rorschachs paired with the corresponding group tests. Again the raters were unaware of whether an individual set of tests had been gathered before or after treatment or whether it belonged to an experimental or control patient.

Comparison Ratings. The next step in the rating procedure consisted of giving each rater a set of 10 pairs of Rorschachs or 10 pairs of group tests, each pair consisting of the pretreatment and posttreatment protocols for a single patient. He was told which of each pair was pre- and which was posttreatment and was requested to complete the Comparison Rating scales. As in the previous rating sessions, each set of 10 contained both experimental and control cases, selected randomly.

Finally, after evaluating several sets of pre-, post-Rorschachs and group tests, each rater was handed a set of 10 complete pairs, each pair consisting of the pretreatment Rorschach and corresponding group tests and the posttreatment Rorschach and group tests of the same patient. The rater was told which were the pre- and which were the posttreatment tests and was requested to complete Comparison Rating scales on the basis of the entire test battery.

Psychotherapy

The experimental group met three times a week for group psychotherapy sessions conducted by the experimenter and a staff psychologist. Each of the therapists met the group alone once a week, and on the third day both acted as cotherapists. In addition, the patients were seen for individual psychotherapy sessions by the experimenter and staff psychologist. They were assigned randomly to each of the psychotherapists for these individual sessions. The time and number of individual sessions per patient were determined in part by the therapist's judgment and in part on the basis of the patients' requests for individual sessions. On the average, two weekly individual meetings were held. All experimental patients were told they could request individual sessions as their needs arose.

The therapists made notes on each patient's behavior immediately following every therapy session (group as well as individual), and each group session was briefly summarized in terms of basic themes and tensions. Many of the sessions, both individual and group, were recorded and filed.⁶

RESULTS AND DISCUSSION

The experimental design of this study was aimed primarily at furnishing at least a partial answer to the question: can the effects of short-term psychotherapy used as an adjunct to insulin-coma treatment be measured over and above the effects of insulin-coma therapy alone? By comparing the changes in relevant

personality variables between pretreatment and posttreatment evaluations of the experimental patients with the changes found in the control patients, it should be possible to determine whether the group receiving psychotherapy in addition to insulin improved significantly more than the group receiving insulin alone.

Analysis of Measures of Personality Used as Criteria for Evaluation of the Effects of Psychotherapy

Two kinds of criteria are available for evaluating the effects of psychotherapy: (a) ratings on a number of personality traits made by clinical psychologists after careful study of individual protocols of psychological tests administered before and after treatment, and (b) changes in ward behavior and psychiatric interview as quantified by scores on the MSRPP.

Psychological Test Ratings. The psychological test scales were analyzed with respect to two main factors: (a) the degree of agreement between two independent raters on each scale, and (b) significance of changes indicated between pre- and posttreatment testing. When the interrater agreement for both scales involved in a pre-post treatment or intergroup comparison was found to be at least .50, the separate ratings of each psychologist were pooled and averaged. When one or both interrater correlations were less than .50, the ratings showing the greatest variability as indicated by the largest standard deviation were used as criterion measures of personality.

The cross-sectional ratings produced lower and more inconsistent reliability than the comparison ratings. In general, the degree of agreement between the two raters was better than chance on both cross-sectional and comparison ratings, although a few correlations were near zero or negative. Comparison of the number of significant correlations between raters judging on the basis of group tests alone, Rorschach alone, or a combination of group tests and Rorschach, revealed little difference. Thus, it seems that increasing the amount of available data on which to base ratings did not increase the degree of agreement between raters. This finding raises an important question for further research, since

⁶ Sample transcriptions may be found in Roos (1955).

many clinicians believe that personality evaluations based on a comprehensive battery of tests are more reliable and valid than those derived from single tests.

Comparison of the mean pre- and posttreatment scores obtained from the experimental and control groups on each of the Cross-Sectional Rating scales revealed marked differences in the ratings and what they signify from one set of tests to the next (e.g., group tests alone, Rorschach alone, and group tests combined with Rorschach). From a total of nine significant (.05 level or beyond) differences between pre- and posttests, eight were derived from ratings based on the group tests alone and one was derived from ratings based on the group tests and Rorschach combined. It appears from these data, therefore, that the group tests were quite sensitive to changes occurring during treatment and that ratings derived from them may be of value for intergroup comparisons. On the other hand, the Rorschach tests—or at least judgments derived from them by the raters in this study—did not significantly reveal these changes. One possible interpretation of these results is that the personality characteristics measured by the Rorschach are “deep” enduring genotypical traits which seemingly remain unchanged as a result of treatment. Another more possible interpretation, however, is that the clinical judgments based on the Rorschach do not reflect the personality characteristics in question.

Comparison ratings on all scales, irrespective of the data upon which they were based, indicated improvement, suggesting that they are suitable for comparing the experimental and control groups. These results seem to conflict with those derived from the cross-sectional ratings which indicated that only those ratings based on the group tests revealed improvement. In all probability, the knowledge of which data were pretreatment and which were gathered after treatment greatly influenced the raters who believed, in general, that most patients were improving. Rater bias should affect the experimental and control groups equally, however, since the raters had no clues as to which protocols were obtained from the experimental group and which were obtained from the control group.

MSRPP Ratings

A measure of psychopathology was derived from the MSRPP by combining the ratings based on the psychiatric interview with the average of the four ratings made by the ward personnel. Comparison of pre- and posttreatment ratings revealed a significant decrease in pathology as measured by the MSRPP (beyond .001 level). These results, however, do not permit any generalizations regarding insulin-coma therapy, since the present study was not designed to evaluate this type of treatment.

Comparison between Experimental and Control Groups

On the basis of the preceding analysis of measures used in this study to assess personality changes, the following types of data were found to reveal changes between pre- and posttreatment evaluations: (a) the cross-sectional ratings based on the group tests, (b) the comparison ratings based on all categories of psychological test data, and (c) the MSRPP measure of pathology derived from psychiatric interview and observed ward behavior. These measures are suitable, therefore, for testing the hypothesis that the experimental group improved more than did the control group.

Psychological Test Ratings. The mean pre-posttreatment changes on the cross-sectional ratings for the experimental and control groups, together with the differences in these changes between the two groups and their statistical significance, are summarized in Table 1. Since both control and experimental patients showed some improvement, only the cross-sectional ratings for which there exist significant differences between the pre- and posttreatment testing for either the control or the experimental group are included in this analysis.

Consideration of these data reveals that all differences in improvement between the experimental and control groups favor the experimental group; that is, the ratings on those scales which differentiated significantly between pre- and posttreatment testing reveal more improvement in the experimental than in the control group. Ratings based on the

TABLE 1

COMPARISON OF EXPERIMENTAL AND CONTROL GROUPS ON CROSS-SECTIONAL RATING SCALES

Rating Scale	Experimental Group Mean Change Score	Control Group Mean Change Score	Difference ^a	<i>t</i>
Group tests:				
Intellectual Efficiency	1.11	0.69	0.42	1.36
Emotional Functioning	1.00	0.25	0.75	2.54*
Psychopathological Characteristics	1.42	0.31	1.11	4.20***
Interpersonal Efficiency—				
Nature of Social Contacts	1.16	0.37	0.78	2.94**
Need Gratification	0.84	0.25	0.59	1.80
Self-Attitudes	0.95	0.63	0.32	1.02
Group tests and Rorschach combined:				
Interpersonal Efficiency—				
Nature of social contacts	0.37	0.13	0.23	0.96

Note.—Only scales where comparison of pre- and posttreatment ratings gave significant *t*'s (beyond .05 level) are presented here.

* Positive values indicate differences favoring experimental group.

* Indicates significance beyond .05 level.

** Indicates significance beyond .01 level.

*** Indicates significance beyond .001 level.

group tests for three scales—Emotional Functioning, Psychopathological Characteristics, and Nature of Social Contacts—reached an acceptable measure of statistical significance (0.05, 0.001, and 0.01 level, respectively). Ratings on Need Gratification approached significance (beyond 0.10 level).

These findings lend strong support to the hypothesis that short-term psychotherapy used as an adjunct to insulin-coma therapy leads to measurable improvement in schizophrenics over and above the changes resulting from insulin-coma treatment alone. If the psychologists' ratings on these three scales actually reflect the traits as defined, the results indicate that short-term psychotherapy helped schizophrenic patients to learn more mature ways of handling their emotional reactions, to abandon primitive psychotic thinking, and to improve their ability to function well in social situations. These are the kinds of changes which most psychotherapists would probably expect to find if psychotherapy were at all successful.

Although none of the differences between the experimental and control groups on the comparison ratings reached statistical signifi-

cance, all the ratings based on the group tests alone favored the experimental group, whereas ratings based on the Rorschach tended to favor neither group consistently.

MSRPP Ratings. Comparisons of the experimental group and the control group with respect to improvement as reflected by changes from the MSRPP revealed that the experimental group showed a greater decrease in pathology than the control group. The *t* test failed to reach statistical significance, however, suggesting there is only a trend favoring the hypothesis that the experimental group made more progress than the control group.

Post-insulin Hospitalization. Since the MSRPP ratings were made shortly following termination of insulin therapy, the temporary leveling effects of the insulin may have overshadowed group differences which would become evident after the immediate effects of the insulin had dissipated. A follow-up study of these and similar cases might reveal more striking differences between the experimental and control groups than those found in this study which compared the groups immediately following treatment. Preliminary follow-up data based on a survey made 2 months

following the present study lend support to this hypothesis. At that time only three of the experimental patients had not been either discharged from the hospital or released on trial visit as compared with eight of the control patients. The average number of days an experimental patient remained hospitalized following the termination of insulin treatment was 38.4 as compared with 72.2 days for the control patient. Since the distribution of numbers of days was markedly skewed, a nonparametric method—the Mann-Whitney U test—was used to determine the significance of the difference between the two groups of patients. A U of 1.94 (p less than 0.05) was obtained, from which it can be concluded that insulin patients receiving psychotherapy are more likely to leave the hospital earlier than are those who receive insulin alone. This finding seems particularly important in view of the fact that the psychiatrist responsible for discharging the patients was unaware of whether any given patient had received psychotherapy.

SUMMARY

The present study was an attempt to test the hypothesis that short-term psychotherapy used as an adjunct to insulin-coma therapy results in measurable improvement in schizophrenics over and above that derived from insulin treatment alone. A control group of 18 schizophrenics undergoing insulin-coma therapy was compared with an experimental group of 19 similar patients who received a combination of group and individual psychotherapy in addition to the insulin treatment.

Evaluative criteria included: (a) pre- and posttreatment ratings by a psychiatrist on the basis of an intensive interview using the MSRPP, (b) biweekly ratings made by ward personnel using the Ward Observation scale of the MSRPP, and (c) pre- and posttreatment ratings made on the basis of individually-administered Rorschachs and a battery of group-administered psychological tests. All ratings were made in such a way as to be free from possible rater bias.

Comparison of the improvement made by the experimental and control groups led to the following major findings: (a) the objective

psychiatric and behavioral ratings revealed differences between the improvement made by the two groups which consistently favored the experimental group, but these trends did not reach statistical significance; (b) the psychological test ratings, which had differentiated between pre- and posttreatment testing, revealed significantly greater improvement in the experimental group than in the control group on three important scales; (c) ratings based on comparisons between pre- and posttreatment protocols failed to differentiate between the mean improvements made by the two groups; and (d) comparison of the two groups on length of time patients remained in the hospital following termination of insulin treatment revealed that the experimental patients—on the average—left the hospital more than a month sooner than the control patients, and this difference proved to be statistically significant. These findings were interpreted as verifying the major hypothesis.

REFERENCES

- FRANK, J. D. Some aspects of lobotomy (prefrontal leucotomy) under psychoanalytic scrutiny. *Psychiatry*, 1950, 13, 35.
- KALINOWSKI, L. B., & HOCH, P. H. *Shock treatments, psychosurgery, and other somatic treatment in psychiatry*. New York: Grune & Stratton, 1952.
- KLOPFER, B., & KELLEY, D. M. *The Rorschach technique*. New York: World Book, 1946.
- LORR, M. Multidimensional scales for rating psychiatric patients, hospital form. *VA tech. Bull.*, 1953, No. TB 10-507.
- LORR, M., JENKINS, R. L., & HOLSOPPLE, J. Q. Factors descriptive of chronic schizophrenics selected for the operation of prefrontal lobotomy. *J. consult. Psychol.*, 1954, 18, 293-296.
- PALMER, D. M., & RIEPENHOFF, T. Insulin shock therapy: A statistical survey of 393 cases. *Amer. J. Psychiat.*, 1950, 106, 918-926.
- ROOS, P. Psychotherapy as an adjunct to insulin-coma therapy in the treatment of schizophrenia. Unpublished doctoral dissertation, University of Texas, 1955.
- ROSEN, J. N. *Direct analysis*. New York: Grune & Stratton, 1953.
- SARGENT, W., & SLATER, E. *An introduction to physical methods of treatment in psychiatry*. Baltimore: Williams & Wilkins, 1948.
- SULLIVAN, H. S. *Conceptions of modern psychiatry*. Washington, D. C.: William Alanson White Psychiatric Foundation, 1940.

(Received August 26, 1960)

A NOTE ON THE SIGNIFICANCE OF DISCREPANCIES BETWEEN GOODENOUGH AND BINET IQ SCORES

ELISE ELKINS LESSING

Illinois Institute for Juvenile Research

Several investigators have tried to ascertain the significance of discrepancies between intellectual ratings derived from the Goodenough Draw-A-Man Test on the one hand and standard intelligence tests on the other. It has been suggested that a negative deviation of the Goodenough IQ from other ratings might be an indication of either brain damage or emotional maladjustment (Goodenough, 1950). Hinrichs (1935) and Brill (1937) found support for this hypothesis in their comparisons of Binet and Goodenough IQ scores in populations of delinquents and defectives, respectively. Hanvik (1953) compared the WISC and the Goodenough IQs of 25 patients in a child guidance clinic. He noted that the mean WISC Full Scale IQ for the group was 13.72 points higher than the mean Goodenough IQ score. In only 4 out of 25 cases did a child's Goodenough IQ exceed his own WISC Full Scale IQ score. He suggested that while the Goodenough might not furnish a valid assessment of the intellectual ability of a clinic population, it might provide an index of neuroticism.

In a pilot study for a larger research project, the present author obtained evidence that the discrepancy between Goodenough IQs and IQs on a standard test of intelligence can be as dramatically large in a nonclinic population as in the disturbed groups studied by Hinrichs and Hanvik. The author had access to male figure drawings made by 21 boys and 2 girls who were described as well-adjusted by their teachers in the Chicago public schools. These children were part of a larger sample used in an Institute study of normal 8 and 9 year old children. One experienced

examiner administered Form L of the Revised Stanford-Binet to all the children and scored the protocols. The same examiner administered the drawing test. The word "person" was substituted for "man" in the Goodenough instructions; however, the encouragement to do the best job possible was retained. The drawings were scored by the author and a trained student worker. A reliability coefficient of .95 was obtained. The student's scoring was used as the basis for all subsequent computations.

The mean Binet IQ of the 23 nonclinic children was found to be 119.96 with an *SD* of 14.84. The mean Goodenough IQ of the group was found to be 92.17 with an *SD* of 17.26. The discrepancy of 27.79 IQ points between the mean Binet and the mean Goodenough IQ of this nonclinic group is actually larger than the discrepancy of 13.72 points which Hanvik found between the WISC and Goodenough IQs of his clinic sample. In the nonclinic group being described, the correlation between the Binet and Goodenough IQ scores was .51, which is significantly different from zero at only the .05 level.

Thus, the situation is somewhat less gratifying than Hanvik's discussion would indicate. The Goodenough has questionable validity as a measure of the intellectual level of both nonclinic and clinic groups. One must be extremely cautious in interpreting a lower Goodenough than WISC or Binet IQ as evidence of emotional disturbance or brain damage: the same pattern of test score relationships can be found in populations presumably characterized by neither of these pathological conditions.

REFERENCES

BRILL, M. A study of instability using the Goodenough drawing scale. *J. abnorm. soc. Psychol.*, 1937, 32, 288-302.

GOODENOUGH, FLORENCE L., & HARRIS, D. Studies in the psychology of children's drawings: II. 1928-1949. *Psychol. Bull.*, 1950, 47, 369-433.

HANVÍK, L. The Goodenough test as a measure of intelligence in child psychiatric patients. *J. clin. Psychol.*, 1953, 9, 71-72.

HINRICHS, W. E. The Goodenough drawing in relation to delinquency and problem behavior. *Arch. Psychol.*, NY, 1935, No. 175.

(Received August 19, 1960)

BRIEF REPORTS

CASE HISTORY DATA AND PSYCHIATRIC DIAGNOSIS¹

EDWARD ZIGLER

Yale University

AND

LESLIE PHILLIPS

Worcester State Hospital

This study investigated how individuals in each of the major groups of functional disorders differed from the population at large and from one another on the variables of age, intelligence, education, occupation, employment history, and marital status. The study was based on an examination of the case history data of 793 patients admitted to Worcester State Hospital during a 12-year period (1945-1957) and referred to the hospital Psychology Department for appraisal. The diagnosis ascribed to each patient was that psychiatric classification agreed upon at a diagnostic staff conference. The patients were categorized into four major diagnostic groups: manic-depressive (37 men, 38 women); schizophrenic (165 men, 122 women); psychoneurotic (81 men, 71 women); and character disorder (197 men, 82 women). The 1950 census data for the state of Massachusetts was employed to compare the total hospital population and the specific diagnostic groups to the population at large on the variables investigated.

An effort was made to determine whether any systematic differences on the biographical variables existed between the population employed in the study and a sample of 111 patients who had not been referred for psychological appraisal. This latter sample was matched with the present

¹ An extended report of this study may be obtained without charge from Leslie Phillips (Worcester State Hospital; Worcester, Massachusetts) or for a fee from the American Documentation Institute. Order Document No. 6829 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance \$2.00 for microfilm or \$3.75 for photocopies. Make checks payable to: Chief, Photoduplication Service, Library of Congress.

This investigation was supported by the Dementia Praecox Research Project, Worcester State Hospital, and a research grant (M896) from the National Institute of Mental Health, United States Public Health Service.

population on the variables of sex, year admitted to the hospital, and diagnosis received. This comparison revealed that there were no differences between the two groups on the variables of occupation, education, or employment history. On the age variable, it was found that the population used in this study is younger than the general hospital sample. In the marital status variable, it was found that patients who are single are equally represented in both samples but that married patients are under-represented while individuals who fall into the "other" category are over-represented in the population used in this study. No comparisons could be made on the intelligence variable. These differences should be kept in mind in generalizing the findings of the present study to a random population of patients suffering from functional disorders.

The most striking finding of this study is that the hospital population and the diagnostic groups that comprise it are not representative of the population at large. In general, hospitalized individuals are drawn from that portion of the population which has the most difficulty in meeting social expectancies.

The results also indicate that the individual diagnostic groups differ from the population at large and from one another in rather specific ways. The distributions of the manic-depressive and psychoneurotic groups on the biographical variables display many similarities, while they differ markedly from the distributions of the schizophrenic and character disorder groups. The distributions of the schizophrenic and character disorder groups also display considerable similarity to one another. In general, the distributions of the manic-depressive and psychoneurotic groups more nearly resemble the distribution of the normal population than do the distributions of the schizophrenic and character disorder groups.

(Received November 23, 1960)

HOMOSEXUAL PREJUDICE AND PERCEPTUAL DEFENSE¹

LOUIS BREGER AND SHEPHARD LIVERANT

Ohio State University

In accordance with Adorno, Frenkel-Brunswik, Levinson, and Sanford's (1950) conceptualization of prejudice (i.e., prejudice serves as a defense against one's own unacceptable impulses) we predicted that individuals scoring high on a measure of homosexual prejudice would show greater indices of threat to homosexual words presented in a perceptual defense situation.

On the basis of their scores on a specially constructed Likert-type scale of manifest attitudes toward homosexuality (H scale) male subjects were divided into a high prejudice group ($N = 21$); a median group ($N = 25$) and a low group ($N = 22$). Following individual administration of the H scale each subject was placed in a perceptual defense situation utilizing the successive carbon method of presentation. The final measure of threat was obtained by comparing the means of each subject on four homosexual, four sexual, and eight neutral words.

Perceptual defense of the avoidance type was demonstrated on the homosexual and sexual words for the total group ($t = 5.65$, $p < .001$ between neutral and homosexual words; $t = 7.50$, $p < .001$ between neutral and sexual words). However, a significant difference was not found between the homosexual and sexual words.

The differential reaction times to the three groups of words by the high, middle, and low H scale scorers indicated that group differences in

defensiveness to the homosexual words of either the avoidance or vigilance type were *not* manifested.

These results consistently fail to support the major hypothesis concerning prejudice and defensiveness. However, assuming the adequacy of our measures, an alternative explanation presents itself. The adequacy of the perceptual defense test as an index of threat tends to be substantiated by the significantly longer reaction times on the taboo words. The high test-retest reliability coefficient (Pearson $r = .91$, $N = 43$), the results of an item analysis, and the successful control of an acquiescence set all indicate that the H scale is measuring some variable in a meaningful manner and a consideration of the undisguised nature of the items makes it appear likely that H scale scores do reflect manifest attitudes toward homosexuality.

The alternative explanation, namely that the attitudes reflected in the H scale are for most subjects learned stereotypes, appears to account for the obtained results. The negative results with the perceptual defense measure are not unexpected, since the notion of learned stereotypes involves no assumptions regarding homosexual prejudice as a defense against repressed impulses. Further corroboration of the stereotype hypothesis is provided by the finding that fraternity members and applied majors score significantly higher on the H scale than nonfraternity men and liberal arts majors, since the former groups would be more likely to hold stereotyped attitudes of all kinds.

REFERENCE

- ADORNO, T. W., FRENKEL-BRUNSWIK, ELSE, LEVINSON, D. J., & SANFORD, R. N. *The authoritarian personality*. New York: Harper, 1950.

(Received November 28, 1960)

¹ An extended report of this study may be obtained without charge from Shephard Liverant (Department of Psychology, Ohio State University; Columbus, Ohio) or for a fee from the American Documentation Institute. Order Document No. 6830 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance \$1.25 for microfilm or \$1.25 for photocopies. Make checks payable to: Chief, Photoduplication Service, Library of Congress.

IMPROVING THE FACTORIAL PURITY OF GUILFORD'S RESTRAINT AND THOUGHTFULNESS SCALES¹

A. W. BENDIG

University of Pittsburgh

Previous factor analytic studies of the 10 scales of the Guilford-Zimmerman Temperament Survey have shown that the Restraint (R) and Thoughtfulness (T) scales have an intercorrelation of approximately .40 and that these scales apparently define one of the four second-order factors in the GZTS: the factor of extraversion-introversion (EI). In order to develop a short and factorially saturated EI scale for personality research the 60 R and T scale items were subjected to four overlapping factor analyses to determine the factor loadings of each item. Within each analysis the items were intercorrelated using the phi coefficient, two centroid factors were extracted using the complete centroid method, and the loadings were analytically rotated to oblique simple structure by the oblimax criterion.² Three groups of college subjects were involved in the analyses: Group I consisted of 300 male freshmen and Group II of 130 male freshmen who completed the full length GZTS inventory, while Group III was composed of 145 subjects of both sexes enrolled in introductory psychology who responded to a subset of R and T items imbedded in another personality inventory.

The R and T scale scores of the Group I subjects were summed and each item correlated with

¹ An extended report of this study may be obtained without charge from A. W. Bendig (Department of Psychology, University of Pittsburgh; Pittsburgh 13, Pennsylvania) or for a fee from the American Documentation Institute. Order Document No. 6833 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance \$1.25 for microfilm or 1.25 for photocopies. Make checks payable to: Chief, Photoduplication Service, Library of Congress.
² The author's appreciation is extended to Gary Otto and William B. Kehl of the university's Computation and Data Processing Center for providing facilities for the statistical analysis.

this EI factor score. The 30 items correlating most highly were administered to Group III and these items factor analyzed. The 60 R and T items were divided into two subsets of 30 items (15 R and 15 T) and two factor analyses performed. The correlations between the R and T scores in both groups were .42, but the average (median) correlation between the two oblique factors was only .17. The 40 items with the most consistent and largest factor loadings were selected for analysis using the item responses of Group II. The correlation between the R and T factors was .21 and only one of the items had loadings inconsistent with the preceding analyses. All four analyses showed certain items to be incorrectly keyed as to factor affiliation.

The responses of the Group III subjects were scored for two pairs of R and T scores: the regular 30-item scales using the Guilford-Zimmerman keying and new 20-item scales using the item keying suggested by the first three factor analyses. The reliabilities of all four scales averaged about .71 (.69 to .73), but the correlation between the 30-item scales of .35 dropped to a nonsignificant .16 between the new 20-item scales.

It was concluded that the moderate correlation between the regular R and T scales was due to the factorially impure item content of the scales and that there is only a small correlation between the R and T factors. It was also suggested that the EI second-order factor may be solely a psychometric artifact of these contaminated scales.

The new 20-item scales were obtained by (a) omitting GZTS Items 2, 12, 17, 42, 47, 62, 92, 107, 112, 122, 153, 163, 168, 178, 208, 213, 223, 228, 243, and 288; (b) Scoring Items 173, 233, and 263 for factor R; and (c) Scoring Items 37, 87, and 97 for factor T. The direction of scoring (true or false) is as given in the GZTS scale keys.

(Received January 3, 1961)

FOOD-RELATED RESPONSES TO AMBIGUOUS STIMULI AS A FUNCTION OF HUNGER AND EGO STRENGTH¹

SEYMOUR EPSTEIN

University of Massachusetts

In a previous study on hunger and thematic apperception (Epstein & Smith, 1956) a theoretical model was proposed which could resolve the discrepancies in a wide variety of studies on the influence of hunger upon food-related responses. According to this model drive can produce an increment, a decrement, or no change in number of drive-related responses. The model integrates Miller's (1951) model of displacement and conflict with the psychoanalytic model of thinking (Rapaport, 1951). Briefly, it is assumed that there are two fundamental processes associated with every drive state: an autistic drive-oriented expressive tendency, and a reality-oriented inhibitory tendency. It is further assumed that the gradient of expression as a function of increasing stimulus-relevance is less steep than the gradient of inhibition. It follows that for stimuli of relatively low relevance, increases in drive should result in an increase in number or intensity of drive-relevant responses, while for stimuli of high relevance the reverse should occur. In this respect, research on both the sex drive (Leiman & Epstein, 1961) and the hunger drive (Epstein & Smith, 1956) have indicated the importance of attending to the drive-relevance of the stimulus. It is further assumed that response-produced cues function in a similar manner to stimulus-produced cues, and that it is as important to consider a dimension of response-relevance as of stimulus-relevance. In this

connection, drive-relevant latent responses, i.e., thoughts and images, are presumed to produce cues which favor inhibition. Finally, it is assumed that there are individual differences in tendency to inhibit drive representatives which can be related to the concept of ego strength.

The purpose of the present study was to investigate different categories of responses, as derived from the above theoretical approach, and to determine whether a measure of ego strength could be related to the influence of drive upon drive-related responses. The Rorschach test was investigated, for, despite the limitation of a small yield of food responses, it offers several scores of interest in terms of the model described, allows these responses to be obtained in a situation where stimulus characteristics play a minimal role, and affords a possible measure of ego strength. The following hypotheses were tested:

1. With increasing hunger there is an increase in food-related responses up to a point, followed by a decrease. This hypothesis follows from the assumption that strong cues, whether stimulus-, response-, or drive-produced favor the inhibitory process. It is consistent with findings in other studies (Levine, Chein, & Murphy, 1942; Wispé, 1954).

2. Food-related responses of low drive-relevance are more strongly associated with hunger than are food-related responses of high drive-relevance. This hypothesis follows from the assumption that responses that produce strong cues are more readily inhibited than responses that produce weak cues.

3. Accurate and popular food responses are more strongly associated with hunger than are inaccurate and unusual food responses. This hypothesis follows from the assumption

¹ This paper was presented, in part, at the Eastern Psychological Association, New York, April 1960. The study is part of a project on the measurement of drive and conflict which is being supported by grant M-1293 from the National Institute of Mental Health, United States Public Health Service. Appreciation is expressed for the assistance of Jane Nelson, Alan Leiman, and Morton Berger.

that the reality-oriented inhibitory process is increasingly dominant at higher drive states, at least up to the point of intense drive at which a breakdown of controls occurs. This view is consistent with findings in a previous study (Levine et al., 1942).

4. Food-related activity-responses are more strongly associated with hunger than food-related object-responses. This hypothesis is based upon the assumption that drive has activating as well as directing properties, and that responses which reflect both are the best drive representatives.

5. People of low ego strength demonstrate a stronger relationship between hunger and food-related responses than people of high ego strength. This hypothesis is based upon the assumption that one of the major aspects of ego strength is the inhibition of drive representatives. It is consistent with reports of marked individual differences in inhibiting thoughts about food (Sanford, 1937).

METHOD

Four levels of hunger were investigated by testing 60 subjects shortly after the noon meal, 60 shortly before the evening meal, 30 before the evening meal after they had abstained from lunch, and 30 before the evening meal after they had abstained from breakfast and lunch. Subjects who missed one meal were paid \$3.00; those who missed two meals \$5.00. The remainder were unpaid volunteers. All were college students, and in each group one-third were female. Subjects were further screened by a questionnaire on what had been eaten when, and by the following self-rating scale on hunger:

Indicate how hungry you feel at the present moment, by placing a check mark to the left of the appropriate statement:

- (a) Not hungry at all (the thought of eating has absolutely no appeal to me at the moment)
- (b) Slightly hungry (would eat something very good, but the thought of food, in general, is not appealing at the moment)
- (c) Fairly hungry (the thought of food is somewhat appealing at the moment, and could enjoy something good)
- (d) Hungry (the thought of food is appealing at the moment, and even something ordinary would be welcome)
- (e) Very hungry (can't wait to eat something; almost anything would taste good)

Finally, groups were equated on total number of Rorschach responses (*R*). The final group consisted of 41 subjects who had not eaten for 0-1 hours with

ratings of a to c on the subjective hunger scale, 40 who had not eaten for 4-5 hours with ratings of c and d, 22 who had not eaten for 8 hours with ratings of d and e, and 21 who had not eaten for 23 hours with ratings of d and e.

Responses were scored in the following categories:

1. Food Imagery—Includes all other categories and consists of all responses with food-association value
2. Strong Food Association—Names of prepared foods, e.g., "fried egg," and people eating or preparing food, e.g., "two people cooking"
3. Weak Food Association—Names of unprepared foods, e.g., "apple"; animals eating or seeking food, e.g., "a squirrel eating a nut"; food-related objects or implements, e.g., "pot," "potato sack"; people in activities of questionable food relevance, e.g., "two people lifting a bowl"
4. Food-Related Object—Names of prepared and unprepared foods and of implements, e.g., "piece of ham," "pot"
5. Food-Related Activity—People or animals seeking, preparing, or eating food, e.g., "two people cooking"
6. Instrumental Responses—People or animals seeking or preparing food; names of foods in inedible form, e.g., "raw egg"; utensils
7. Goal Responses—Food in edible form; people or animals eating
8. Accurate Food—Food, prepared or unprepared, which accurately corresponds to the contours of the blot
9. Inaccurate Food—Food, prepared or unprepared, which does not accurately fit the blot
10. Popular Food—Prepared or unprepared food produced at least six times to the same blot area in the total sample
11. Original Food—Prepared or unprepared food produced no more than once to a particular blot area

Finally, "ego strength" was measured by the form-level score of Klopfer (Klopfer, Ainsworth, Klopfer, & Holt, 1954), with the exception that total rather than average form-level was used, as essentially what was wanted was a score of goodness of response, and it was assumed that, holding quality constant, producing more responses indicated more ability than producing fewer responses. Although total form-level was directly related to *R* and might be expected to be directly related to number of food responses, the relationship was actually inverse and did not approach significance. Thus there was no basis for concern over the two measures being confounded. In defense of the measure of ego strength, it includes perceptual accuracy, integrative ability, and self-imposed motivation, all of which are characteristics of ego strength.

Before scoring, the data were coded to prevent scoring bias. In addition, the data on food-relevant responses and on ego-strength were separately represented and independently scored to prevent confounding of the measures derived from them.

RESULTS

Comparisons were made of the number of subjects in each group who fell above and below the median cutting point for the pooled group. In each category the cutting point turned out to be between zero and one response. Two-tailed chi square tests were used to evaluate significance.

In Figure 1 it can be seen that self-rated hunger increases to 8 hours of deprivation and then levels off. The results are essentially the same when the entire pool of subjects is used as when the sample is restricted to subjects screened on subjective hunger and matched on *R*. The similarity of the subjective hunger ratings of the 8- and 23-hour groups suggests that hunger may fail to increase between 8 and 23 hours of deprivation and raises the question of whether the two groups should be treated separately or combined to provide a sample as large as for the other groups. Accordingly, where indicated, the data were analyzed both ways.

In regard to Hypothesis 1, which stated that there is an increase in food imagery up to a point and then a decrease, Figure 2 indicates that this, in fact, did occur. The relationship between total food imagery and time without food, however, is not statistically significant ($p = .15$). If weak food associations are substituted for total food imagery, a falling off of responses again is indicated (see Figure 2), but the relationship now becomes significant (.05 level). Apparently strong food associations reduce the discriminability of the total food imagery score. When the 8- and 23-hour groups were com-

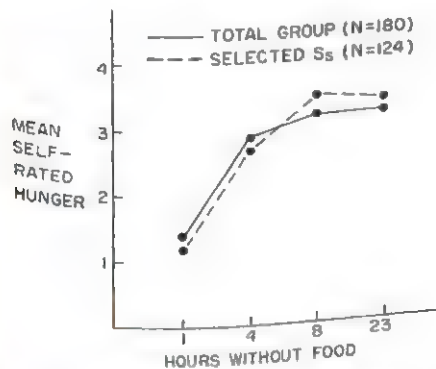


FIG. 1. Self-rated hunger as a function of time without food.

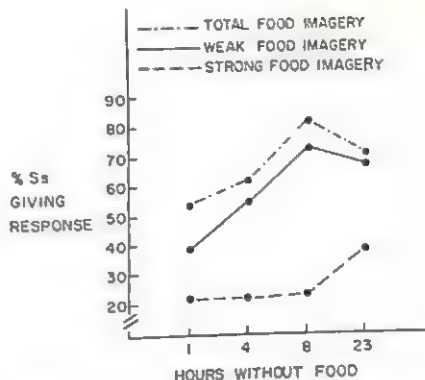


FIG. 2. Total, weak, and strong food associations as a function of time without food.

bined, both total food imagery (.05 level) and weak food imagery (.01 level) increase directly and significantly with increasing hunger. It may be concluded that as deprivation increases there is an increase in weak and overall food associations, up to a point, followed by a levelling off, or possibly decrease, somewhere between 8 and 23 hours of deprivation.

The finding of a significant relationship between weak, but not strong, food associations and time without food is consistent with Hypothesis 2, which stated that responses of low drive-relevance are more strongly associated with hunger than responses of high drive-relevance. Despite this, the results do not entirely support the model, as according to the model strong food associations should fall off more rapidly than weak food associations, whereas Figure 2 indicates that the reverse tended to occur.

The only additional score which significantly (.05 level) discriminated the hunger groups was the food activity score, which, in line with hypothesis, varied directly with hunger.

In order to investigate the effects of ego strength, form-level scores were summed for all nonfood responses, and a cutting point selected to divide the group as nearly in half as possible. This resulted in 67 subjects in the low form-level group and 57 in the high form-level group. In Figure 3 it can be seen that for the low ego strength group there is an increase in total food imagery from 1 to 23 hours, whereas for the high ego strength

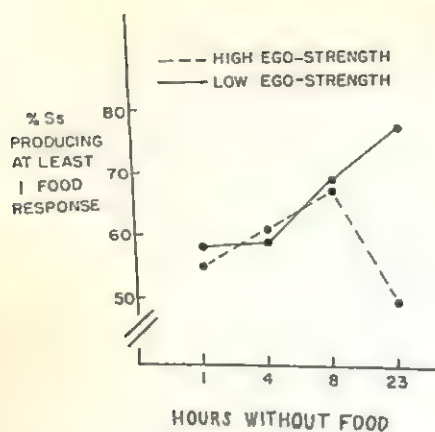


FIG. 3. The influence of ego strength upon the relationship of food associations to time without food.

group there is an increase followed by a decrease.

A comparison of the high and low ego strength groups on all levels of deprivation, simultaneously, fails to reveal significant differences. Significance (.05 level) is obtained, however, if the comparison is restricted to the 23-hour groups. This latter finding must be treated with caution, as it could be a consequence of capitalizing on chance probabilities.

In order to determine whether the decrease in food responses for the high ego strength group reflects a decrease in drive rather than an increase in response-inhibition, hunger ratings were examined in reference to ego strength. The high and low ego strength groups both reported an increase in hunger up to 8 hours and a leveling off to 23 hours. The groups did not differ at 1 and 4 hours, but at 8 and 23 hours the high ego strength subjects rated themselves significantly (.05 level) less hungry than the low ego strength subjects (chi squares were, respectively, 4.70 and 4.87, $df = 1$). The self-ratings suggest that the high ego strength group either inhibits awareness of hunger or is actually in a state of reduced hunger as compared to the low ego strength group.

DISCUSSION

It was found that with increasing time without food there was an increase in food-related responses followed by a decrease. However, the relationship was significant only when strong food associations were eliminated.

Taken in conjunction with findings in other studies (Clarke & Epstein, 1957; Lazarus, Yousem, & Arenberg, 1953; Levine et al., 1942; McClelland & Atkinson, 1948; Sanford, 1937; Wispé, 1954; Wispé & Drambarean, 1953), it would seem safe to conclude that with increasing time without food there is an increase followed by a leveling off or decrease in at least certain types of food response. This leveling off or decrease has sometimes been interpreted to indicate the operation of response-inhibition at higher drive levels. There would be a serious difficulty with such an interpretation in the present study as the 23-hour deprivation group rated themselves as no more hungry than the 8-hour group. Moreover, subjective estimates of hunger at 23 hours were more likely overestimates than underestimates, as they were probably influenced by awareness of length of deprivation. Thus, the decrease in food responses at 23 hours may better be interpreted as reflecting a leveling or falling off in hunger rather than an inhibition of drive-related responses, i.e., as indicating drive-inhibition rather than response-inhibition.

Some evidence possibly supporting response-inhibition was presented by the finding that weak food associations were significantly associated with time without food while strong food associations, which are more susceptible to inhibition, were not. However, according to the model the strong associations should have fallen off more rapidly with increasing hunger than the weak food associations, whereas the reverse tended to occur. Thus, if the basic model is to be preserved, it will be necessary to give further consideration to the complicating effects of response-inhibition.

The findings on ego strength offer some interesting evidence for individual differences in inhibition. For the groups of low ego strength, a direct relationship was found between food-related responses and deprivation through 23 hours of deprivation. For the groups of high ego strength, a direct relation was found through 8 hours of deprivation, but the 23-hour group produced significantly fewer food-related responses than the 8-hour group. The low and high ego strength groups both produced negatively accelerated curves for self-

rated hunger as a function of deprivation. However, the 8-hour and 23-hour groups of high ego strength rated themselves as significantly less hungry than the corresponding groups of low ego strength. The combined evidence suggests, in accordance with hypothesis, that high ego strength subjects are more apt to inhibit than low ego strength subjects. If the self-ratings are accepted as a true report of hunger, the extremely low number of food-related responses produced by the high ego strength group at 23 hours of deprivation supports the occurrence of response-inhibition, i.e., holding subjective hunger constant, high ego strength subjects are less apt to give drive-related responses than low ego strength subjects.

Very likely inhibition of food-related imagery and responses serve to reduce drive, so that the two types of inhibition are not unrelated. That response-inhibition need not be a conscious process was indicated by the questionnaire at the end of the study in response to which almost all subjects denied suppressing food-responses.

Apart from the model proposed, the finding that weak food associations varied significantly with hunger while strong ones did not is of considerable interest. Coupled with negative findings on a difference between instrumental and goal responses, it suggests that a dimension of response-relevance is more fundamental than an instrumental-goal distinction. The reports in other studies (Atkinson & McClelland, 1948; McClelland & Atkinson, 1948; Wispé, 1954; Wispé & Drambarean, 1953) that instrumental responses provide better indices of drive than goal responses can be explained by considering that instrumental responses are generally of lower drive-relevance than goal responses. Moreover, grouping responses of food in an inedible state (e.g., "wheat") with food-related implements (e.g., "table") because they are presumably both "instrumental" to eating would seem more forced than classifying them as of relatively low food-relevance.

In line with hypothesis, it was found that food-related activity-responses were directly and significantly associated with hunger while food-related object-responses were not. The hypothesis was based upon the consideration

that drive has both directing and activating aspects, and that associations that reflect both are better drive representatives than associations that reflect only the directive aspect.

No support was found for the hypothesis that popular and accurate food-related responses increase more regularly with hunger than unusual and inaccurate food-related responses. The hypothesis was based on the assumption that a reality-oriented inhibitory process is dominant at higher drive states, at least up to a point of breakdown. Levine et al. (1942) concluded that as drive increases, the organism becomes increasingly realistic in responding to drive-relevant stimuli. McClelland (1951) takes much the same view in hypothesizing that with increasing deprivation a "reality stage" follows a "wish fulfillment stage" which only under intense deprivation is superceded by a "defense stage" where wish fulfillment again becomes dominant. All that can be said at this point is that the data are inconclusive, and that the hypothesis of increasing accuracy of drive-related responses with increasing drive up to some limit, although reasonable, has yet to be experimentally confirmed. The evidence provided by Levine et al. (1942) is particularly open to question, as it is based on repeated testing of five subjects without control for practice effects, and the explanation was *a posteriori*.

A serious limitation in the present study was the number and quality of food-related responses elicited by the Rorschach test. When this is considered together with the number of comparisons made, and taken in conjunction with evidence that set-effects in laboratory studies are apt to be more important than drive-effects (Clarke & Epstein, 1957; Postman & Crutchfield, 1952; Taylor, 1956), the need for replication under varied conditions is clearly indicated. That factors other than hunger were complicating the food-related responses was indicated by the bizarre nature of some of the responses, e.g., "Dante's inferno, the bottom is the fiery tombs of the heretics, on the sides are the pigs of the gluttons, and at the top are the mournful souls who lived too early, the virtuous pagans." In laboratory investigations on the directive influence of drive, a major diffi-

culty is that the magnitude of the drive is relatively small in comparison with other effects. Requiring the subject to abstain from eating itself introduces set-effects, and informing all subjects that the study is food-related offers only a partial solution, as the information supported by abstinence has a different effect from the same information not so supported. The only solution to this problem is either to investigate no more than 4 or 5 hours of deprivation, or to obtain subjects who have not eaten for reasons unconnected with the study. Fortunately, there were several subjects who were assigned to the control group but were eliminated because they had missed one or more meals. Follow-up interviews were held with seven such subjects who had not eaten for 8-23 hours and who rated themselves as "hungry" to "very hungry." Not one of these produced a strong food association and five produced weak food associations. The results on this group are consistent with those on the larger experimental group, and suggest that set-effects do not explain away the findings on association strength nor the finding of a negative acceleration of food responses as a function of deprivation.

SUMMARY

The present study was undertaken to investigate different categories of food responses to ambiguous stimuli as influenced by hunger and ego strength. Subjects consisted of male and female college students divided into four levels of deprivation: 41 who had not eaten for 0-1 hours, 40 for 4-5 hours, 22 for 8 hours, and 21 for 23 hours. Food-related scores and a measure of ego strength were obtained from a Rorschach test.

The major findings were as follows:

1. Subjective hunger ratings as a function of time without food increased through 8 hours, but did not increase further at 23 hours. This was interpreted as suggesting that hunger itself may level off or decrease somewhere between 8 and 23 hours of deprivation, and that drive-inhibition probably occurs.

2. Overall food imagery increased through 8 hours of deprivation and decreased at 23 hours. However, the relationship reached sta-

tistical significance only when strong food associations were eliminated. This was interpreted as supporting drive-inhibition and indicating that strong food associations, since they are more easily inhibited, are more susceptible than weak food associations to influence by factors other than drive.

3. A group of high ego strength subjects reported significantly less hunger at 8 and 23 hours of deprivation, and produced significantly fewer food-related responses at 23 hours of deprivation, than a group of low ego strength subjects. Only the high ego strength group demonstrated a decrease in food responses at 23 hours. It was concluded that ego strength is related to the inhibition of both drive and drive-related responses.

4. Food-related activity-responses were significantly and positively related to deprivation; food-object responses were not.

5. Significance was not found for goal and instrumental food responses. It was proposed that the goal-instrumental distinction could be subsumed under a dimension of drive-relevance of the response.

6. There was no evidence that food responses become more accurate or stimulus-determined as deprivation increases.

REFERENCES

- ATKINSON, J. W., & McCLELLAND, D. C. The projective expression of needs: II. The effect of different intensities of the hunger drive on thematic apperception. *J. exp. Psychol.*, 1948, **38**, 643-658.
- CLARKE, A. R., & EPSTEIN, S. Food-related responses to ambiguous stimuli as a function of time without food and experimental set. *Amer. Psychologist*, 1957, **12**, 394. (Abstract)
- EPSTEIN, S., & SMITH, R. Thematic apperception as a measure of the hunger drive. *J. proj. Tech.*, 1956, **20**, 372-384.
- KLOPFER, B., AINSWORTH, MARY D., KLOPFER, W. G., & HOLT, R. R. *Developments in the Rorschach technique*. Vol. 1. Yonkers, N. Y.: World Book, 1954.
- LAZARUS, R. S., YOUSEM, H., & ARENBERG, A. Hunger and perception. *J. Pers.*, 1953, **21**, 312-328.
- LEIDMAN, A., & EPSTEIN, S. Thematic sexual responses as related to sexual drive and guilt. *J. abnorm. soc. Psychol.*, 1961, **63**, 169-175.
- LEVINE, R., CHEIN, I., & MURPHY, G. The relation of the intensity of a need to the amount of perceptual distortion. *J. Psychol.*, 1942, **13**, 283-293.
- McCLELLAND, D. C. *Personality*. New York: Dryden, 1951.

- McClelland, D. C., & Atkinson, J. W. The projective expression of needs: I. The effect of different intensities of the hunger drive on perception. *J. Psychol.*, 1948, 25, 205-222.
- Miller, N. E. Comments on theoretical models illustrated by the development of a theory of conflict behavior. *J. Pers.*, 1951, 20, 82-100.
- Postman, L., & Crutchfield, R. S. The interaction of need, set, and stimulus-structure in a cognitive task. *Amer. J. Psychol.*, 1952, 65, 196-217.
- Rapaport, D. The conceptual model of psychoanalysis. *J. Pers.*, 1951, 20, 56-81.
- Sanford, R. N. The effect of abstinence from food upon imaginal processes: A further experiment. *J. Psychol.*, 1937, 3, 145-159.
- Taylor, J. A. Physiological need, set, and visual duration threshold. *J. abnorm. soc. Psychol.*, 1956, 52, 96-99.
- Wispé, L. G. Physiological need, verbal frequency, and word association. *J. abnorm. soc. Psychol.*, 1954, 49, 229-234.
- Wispé, L. G., & Drambarean, N. C. Physiological need, verbal frequency, and visual duration thresholds. *J. exp. Psychol.*, 1953, 46, 25-31.

(Received October 12, 1960)

THE RELATIONSHIP BETWEEN FUTURE TIME PERSPECTIVE, TIME ESTIMATION, AND IMPULSE CONTROL IN A GROUP OF YOUNG OFFENDERS AND IN A CONTROL GROUP

ARON WOLFE SIEGMAN¹

University of Maryland School of Medicine

Recent years have witnessed an increasing interest in time as a psychological variable. There has been concern with sources of variance in time estimation and with individual differences in time orientation (Wallace & Rabin, 1960). The present study attempts to relate both of these dimensions. More specifically, this study investigated the hypothesis that the range of one's future time perspective is a significant source of variance in the experience of duration: the longer the range of one's future time perspective, the faster one's internal clock. This hypothesis is based on the following assumptions: that the more the subject desires that an interval of time pass rapidly, the longer it will appear to be (Irwin, 1961, p. 235), and that subjects with a long range future time perspective are relatively more motivated that time pass rapidly. A positive correlation is thus predicted between future time perspective and time estimation scores. There are also some empirical data which suggest the hypothesized relationship between future time perspective and the experience of duration. Hindle (1951) and Meade (1959) found that when engaged in a task, the subject's estimation of elapsed time is a function of the distance of the goal of the task. The further the goal the greater the subject's estimation of the elapsed time. This finding is consistent with a motivational approach to time estimation which holds that

the more the subject desires that an interval of time pass rapidly, the longer it will appear to be. Generalizing from this finding one can predict a positive correlation between future time perspective—defined as the relative distance of one's life goals—and time estimation. A study by Knapp and Garbutt (1958) on the relation between achievement motivation and time imagery also suggests the hypothesized relationship between future time perspective and the experience of duration. In this study it was found that subjects with high achievement motivation described time in metaphors which reflected a very rapid internal clock. Since there is evidence which indicates that need achievement is one of the sources of future time perspective (Siegman, in press-b), the findings of Knapp and Garbutt (1958) suggest a positive correlation between the range of subjects' future time perspective and the speed of subjects' internal clock.

The hypothesized positive correlation between future time perspective and time estimation scores generates the additional prediction that delinquents will have shorter time estimation scores than nondelinquents. This prediction is based on the finding (Barndt & Johnson, 1955) that delinquents have significantly shorter future time perspectives than comparable nondelinquents.

A second objective of this study was to investigate the relationship between impulse control and future time perspective. LeShan (1952), in one of the first empirical investigations of future time perspective, suggests the hypothesis that the more thoroughly a child learns to delay the immediate gratification of his impulses for the sake of some other

¹ This study was conducted while the author was at Bar-Ilan University, Israel. The author is grateful to A. Nir, Commissioner of Israel's Prison Service; to T. Givati, Director of the Tel-Mond Prison; and to E. Carni of the Israel Army for their help in the procurement of subjects. The author is also indebted to Jacob Jonah for his assistance in the collection of the data for this study.

and more distant goal, the greater his future time perspective as an adult. Assuming that there is less learning of such control among lower-class than middle-class children, LeShan (1952) predicted and found significantly shorter future time perspectives among lower-class children. Furthermore, assuming that delinquents have failed to acquire such delay capacity, LeShan (1952) argues that they should also have relatively more restricted future time perspectives. This prediction was confirmed in a subsequent investigation (Barndt & Johnson, 1955). These findings, however, cannot be considered as sufficient evidence for the hypothesis that future time perspective is a function of impulse control training. It is clear that delinquents and nondelinquents, and lower and middle-class children differ from each other in relation to many other variables than impulse control training, some of which may be responsible for the differences obtained in relation to future time perspective. Consequently, the present study investigated the hypothesized relationship between future time perspective and impulse control training in a more direct fashion. Actually, this study investigated the relation between subjects' future time perspective and subjects' present impulse control level.

SUBJECTS AND PROCEDURE

The delinquent group consisted of 30 residents at the Tel-Mond Prison for Young Offenders, Israel. Subjects were selected according to alphabetical order from the age range 17-19. The education of this group ranged from 1 to 8 years, with Mean = 5.9 and $SD = 1.28$. The nondelinquent group consisted of 22 subjects who, in order to control for institutionalization, were selected from among recent inductees in the Israeli Army. Subjects for the nondelinquent group were selected so as to obtain an age and educational distribution identical to the experimental group. The two groups were also equated for ethnic origin, with 77% of Middle-Eastern or North African origin and 23% of European origin. Subjects of both groups were of lower socioeconomic background.

Subjects' future time perspectives were determined by a method similar to that used by Wallace (1956). Each subject was asked to enumerate 10 events that refer to things which he may do or which may happen to him in the future. At the completion of this task, the subject was asked to indicate what age he thought he would be at the occurrence of each of the events. The median of the differences between the subject's present age and the ages indicated by the

TABLE 1
FUTURE TIME PERSPECTIVE MEAN AND STANDARD
DEVIATION SCORES IN DELINQUENT AND
NONDELINQUENT GROUPS

Group	N	M	SD
Delinquent	30	3.10	.412
Nondelinquent	22	4.95	.411

subject for the various events was used as the subject's future time perspective score.

In the time estimation task, each subject was presented with the following time intervals: 5, 25, and 15 seconds. The beginning and end of each interval was marked by the click of a stop watch. The intervals between stimuli were 5 seconds. In order to control for the serial position effect found by previous investigators (Eson & Kafka, 1952; Falk & Bindra, 1954; Siegman, in press-a), one half of each group was presented with the stimuli in the order in which they are listed, and the other half in the reversed order. All subjects were told that the purpose of this study was to determine how they experienced various periods of time, not to count off the seconds, and not to make use of mnemonic devices. Two subjects of the nondelinquent group did not participate in the time estimation task.

Subjects' impulse control level was determined by means of a motor inhibition task. In this task, subjects were instructed to trace a 2½-inch circle on onion skin paper as slowly as possible. This task is a variation on a task which was used by Singer, Wilensky, and McCraven (1956), who asked subjects to write certain words as slowly as possible. A factor analytic study, and a number of other studies in which this task was used as a measure of subjects' impulse control level, provide considerable construct validity for this kind of task (Singer et al., 1956).

Progressive Matrices (PM) scores were available for all subjects of the control group.

RESULTS

Table 1 indicates that the delinquent group obtained, as was hypothesized, lower future time perspective scores than the nondelinquent group. Since subjects' future time perspective scores were not normally distributed, the significance of the difference between the two groups was evaluated by the Mann-Whitney *U* test (Siegel, 1956, pp. 116-127). The results were: $U = 130.5$, $p < .0003$.

Table 2 lists the time estimations of the delinquent and nondelinquent groups. As predicted, the delinquent group obtained the lower time estimation scores. Because of heterogeneity of variance and because the time

available empirical evidence is equivocal (Siegman, 1961).

Since subjects' motor impulse control scores were not normally distributed, the significance of the difference between the two groups was determined by the Mann-Whitney U test. The results were: $U = 305$, which is clearly not significant ($p = .65$). This finding suggests that delinquents are not, as is generally assumed, less able to control their impulses. The common observation that delinquents have a history of impulsive behavior may be due to the fact that they are less motivated to control their impulses rather than to defective control mechanisms. The fact that most delinquents come from a socioeconomic background which does not provide them with sufficient incentives for controlling their anti-social impulses (Cohen, 1955) is perhaps one factor responsible for this lack of motivation.

Finally, the data of the present study make it possible to evaluate the relationship between impulsivity and the experience of duration. In a previous investigation (Siegman, in press-a) a negative correlation was obtained between subjects' scores on a motor impulse inhibition task and subjects' verbal estimations of brief time intervals. This finding suggests that impulsive subjects have relatively more rapid internal clocks. The correlation failed, however, to reach the conventional .05 significance level. In the present study the correlations (Kendall's tau) between subjects' scores on the motor impulse control task and their estimations of the 5-, 15-, and 25-second time intervals were $-.28$ ($p = .08$), $-.29$ ($p = .07$), and $-.27$ ($p = .09$) in the nondelinquent group, and $-.21$ ($p = .10$), $-.22$ ($p = .08$), and $-.21$ ($p = .10$) in the delinquent group. Again the correlations fail to reach the .05 level of significance. The fact, however, that the correlations vary consistently between the .05 and .10 level of significance is suggestive of the hypothesized relationship between impulsivity and the experience of duration.

SUMMARY

A positive correlation was found in a group of young delinquents and in a comparable group of nondelinquents, between the range of subjects' future time perspective and the

speed of subjects' internal clock as measured by a time estimation task.

The delinquent group obtained significantly lower future time perspective scores as well as lower time estimation scores.

A significant positive correlation was obtained, in the delinquent group, between subjects' future time perspective scores and their scores on a motor impulse control task. The correlation between these two variables in the nondelinquent group, however, was not significant.

There was no significant difference between the delinquent and nondelinquent group in relation to their scores on the motor impulse inhibition task.

Negative correlations were obtained between subjects' time estimation and motor impulse inhibition scores, which were significant between the .05 and .10 level.

Finally, the results of the present study also suggest that there is no significant correlation between general intelligence and future time perspective or impulse control.

REFERENCES

- BARNDT, R. J., & JOHNSON, D. M. Time orientation in delinquents. *J. abnorm. soc. Psychol.*, 1955, 51, 343-345.
- COHEN, A. K. *Delinquent boys: The culture of the gang*. Glencoe, Ill.: Free Press, 1955.
- ESON, M. E., & KAFKA, J. S. Diagnostic implications of a study in time perception. *J. gen. Psychol.*, 1952, 46, 169-183.
- FALK, J. L., & BINDRA, D. Judgment of time as a function of serial position and stress. *J. exp. Psychol.*, 1954, 47, 279-282.
- GLUECK, S., & GLUECK, E. *Unraveling juvenile delinquency*. New York: Commonwealth Fund, 1950.
- GREENACRE, P. Conscience in the psychopath. *Amer. J. Orthopsychiat.*, 1945, 15, 495-509.
- HINDLE, H. M. Time estimates as a function of distance travelled and relative clarity of a goal. *J. Pers.*, 1951, 19, 485-501.
- IRWIN, F. W. Motivation and performance. *Annu. Rev. Psychol.*, 1961, 12, 217-242.
- KNAPP, R. H., & GARBUIT, J. T. Time imagery and the achievement motive. *J. Pers.*, 1958, 26, 426-434.
- LESHAN, L. L. Time orientation and social class. *J. abnorm. soc. Psychol.*, 1952, 47, 582-589.
- LEVINE, M., GLASS, H., & MELTZOFF, J. The inhibition process, Rorschach human movement responses, and intelligence. *J. consult. Psychol.*, 1957, 21, 41-45.
- MEADE, D. R. Time estimates as affected by motivational level, goal distance, and rate of progress. *J. exp. Psychol.*, 1959, 58, 275-279.

- MICHAELS, J., & STEINBERG, A. Persistent enuresis and delinquency. *Brit. J. Delinqu.*, 1952, 3, 1-10.
- RAPPAPORT, D. *Organization and pathology of thought*. New York: Columbia Univer. Press, 1951.
- SIEGEL, S. *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill, 1956.
- SIEGMAN, A. W. Arahei hahevra vchitnahagut hayahid. [Cultural values and human behavior.] In, *Baayot calijot bahinuch haisraeli*. Jerusalem: Ministry of Culture and Education, 1959. Pp. 17-37.
- SIEGMAN, A. W. Theories of juvenile delinquency: Some empirical investigations. Paper read at fourteenth International Congress of Applied Psychology, Copenhagen, August 1961.
- SIEGMAN, A. W. Anxiety, impulse control, intelligence, and the estimation of time. *J. clin. Psychol.*, in press. (a)
- SIEGMAN, A. W. Some variables associated with future-time perspective. *Darshana*, in press. (b)
- SINGER, J. L., WILENSKY, H., & MCCRAVEN, V. G. Delaying capacity, fantasy, and planning ability: A factorial study of some basic ego functions. *J. consult. Psychol.*, 1956, 20, 375-383.
- SPIVACK, G., LEVINE, M., & SPRINGLE, H. Intelligence test performance and the delay function of the ego. *J. consult. Psychol.*, 1959, 23, 428-431.
- TEAHAN, J. E. Future time perspective, optimism, and academic achievement. *J. abnorm. soc. Psychol.*, 1958, 57, 379-381.
- WALLACE, M. Future time perspective in schizophrenia. *J. abnorm. soc. Psychol.*, 1956, 52, 240-245.
- WALLACE, M., & RABIN, A. I. Temporal experience. *Psychol. Bull.*, 1960, 57, 213-237.

(Received April 21, 1960)

GROUP PSYCHOTHERAPY, A SPECIAL ACTIVITY PROGRAM, AND GROUP STRUCTURE IN THE TREATMENT OF CHRONIC SCHIZOPHRENICS¹

JAMES M. ANKER AND RICHARD P. WALSH²

Veterans Administration Hospital, Perry Point, Maryland

Activity programs and group psychotherapy frequently have been used in the treatment of the chronic patient. Evaluations of these procedures, unfortunately, often have suffered from poor definition and choice of procedure, and inadequate design. This paper reports an experimental study of these therapies in combination with the effect of group homogeneity.

The use of activity programs in neuropsychiatric hospitals received its major impetus from Myerson's "total push" method in the treatment of chronic NP hospital patients (Myerson, 1939). Many authors since have reported varying degrees of success with this treatment method and progressive variations upon it (e.g., Cohen, 1959; Murray & Cohen, 1959; Pace, 1957; Peters, 1955). Patients in such programs usually have been assigned to a group and given or offered a relatively specific "job." Responsibility for performance typically rests with the supervisor or therapist and slightly, if at all, with the patient members. Generally this approach might be thought of as a systematized attempt to increase drive level by stimulation and encouragement.

Scher (1957a, 1957b) extended this concept considerably with his emphasis upon the "task orientation" of the patient. When response to the task was demanded rather than

being requested Scher concluded that significant therapeutic changes were produced. A controlled study by DiGiovanni (1958) failed to replicate Scher's results when the activity group was compared with a psychotherapy group and a control group, and all groups were compared before and after. The treatment procedures were conducted for only 4 months, however, as compared with 12 months in Scher's study.

Members of groups occurring naturally adopt responsibility and/or develop ways of delegating it to themselves and other members. They participate in a general division of labor. Kretch and Crutchfield (1948) state this mutual cohesiveness, as opposed to external pressure, is what constitutes a group. It was presumed that this type of social organization could occur in a chronic schizophrenic group and, to the extent that it did occur, behavior alteration would be possible. This kind of group may be distinguished from many extant hospital activity groups by the locus of responsibility, the patient members themselves.

The activity program evaluated in this study was designed to promote this "normal" type of social organization. The criteria chosen for it were as follows: (a) the group should have a definite goal or finished product which may be achieved in a relatively short period of time; (b) the goal should be periodic so that once the immediate goal is achieved another similar one, but one presenting new challenges, takes its place; (c) there must be a sufficient range of demand so that patients at all levels of adjustment may contribute meaningfully to the goal; (d) the activity must be complex enough so that it will pose meaningful problems to be resolved by the

¹ A review of this paper was presented at the fifth annual Research Conference, Cooperative Chemotherapy Studies in Psychiatry and Research Approaches to Mental Illness, Cincinnati, June 6-8, 1960, and at the eleventh semiannual Veterans Administration-Universities Conference, Washington, D. C., December 1, 1960.

² Formerly counseling psychologist, Psychology Service, Veterans Administration Hospital, Perry Point, Maryland. Now Chief Counselor and Assistant Professor in Psychology, Testing and Counseling Center, University of Cincinnati.

group; and (e) this activity must be of such a nature that patients are capable of maintaining it with a minimum of staff intervention, particularly professional staff. After evaluating a number of possible programs, the activity chosen for study was the production of plays for hospital patients and personnel. The characteristics of this activity are described in more detail under "Method."

It would be insufficient to describe the type of group psychotherapy studied as "orthodox." A review of pertinent literature in the area reveals a rather widespread range of approaches to group psychotherapy with schizophrenics (e.g., Bach, 1954; Grauer, 1955; Klapman, 1946, 1947; Kramer, 1957; Lazell, 1945; Peyman, 1956; Powdermaker & Frank, 1953; Schnadt, 1955). The technique used in this study very closely resembles that described by Kramer with the possible exception of less emphasis being placed upon the role of interpretation by the therapist. The atmosphere was permissive and designed to promote a growing sense of "belongingness" by fostering a comfortable "family quality." During the sessions primary emphasis was placed on nonpsychotic interactions between patients. Interaction was encouraged and implemented by the therapist but not demanded. The therapist patterned his orientation after that described by Frank (1952) by being "a perspicacious, strong, accepting person who structured the situation clearly for the patients and supported them in their emotional turmoil." Any level of nonpsychotic verbal interaction was encouraged. This included topics like the difficulty in keeping personal belongings identified in the hospital laundry and the problem of saving enough money for passes. This technique might be contrasted to the didactic or pedagogical approach suggested by Lazell and Klapman.

A number of authors (Hoffman, 1959; Kramer, 1957; Powdermaker & Frank, 1953), most writing on group psychotherapy, have speculated on the advantages of group homogeneity or heterogeneity. While there is active disagreement in this area, the consensus favors some type of heterogeneity. Group structure, because of its implications for group treatment methods generally, was included as a main effect in this study.

It was hypothesized that significant improvement in behavioral adjustment would occur as a result of group psychotherapy, the special activity program (drama group), and heterogeneous group structure. The analysis of these independent variables and of their interactions constitute the study reported here.

DESIGN OF STUDY

The three independent variables and their interactions were analyzed simultaneously by a $2 \times 2 \times 2$ factorial design, each variable being dichotomized. The effectiveness of group therapy was evaluated by contrast with a comparable group not receiving group therapy, the effectiveness of the drama group by contrast with a comparable group not in the special activity program, and the effectiveness of heterogeneity by contrast with a comparable homogeneous group. Because a patient's original level of behavioral adjustment could influence the degree of change in adjustment the data were adjusted for this effect by covariance. Additionally, it was impossible to insure that all patients would remain in the study until its conclusion. Because the length of exposure to the treatment procedures could effect the degree of change in adjustive behavior this effect was covaried as well. Thus the design was a 2^3 factorial analysis of multiple covariance. The unique combinations of the three dichotomous variables resulted in eight distinct "treatment" groups.

PROCEDURE

Selection of Subjects and Groups

One-hundred-thirty-four male schizophrenic patients on a continued treatment ward of a 1,500 bed Veterans Administration Neuropsychiatric Hospital were rated on the Multidimensional Scale for Rating Psychiatric Patients (Lorr, 1953). A pilot study of interrater reliabilities produced an average reliability coefficient of .85 taken over 11 ward personnel. The average interrater reliability coefficient for three raters on the interview section was .91. Coefficients in the total matrix ranged from .66 to .96. This level of reliability was considered sufficient to allow ratings by different raters to be considered as comparable. The protocols were scored for each patient and each profile was compared with the hypothetical normal profile by means of the *D* statistic (Osgood & Suci, 1952). A distribution of *D*s thus was generated, one end of the distribution reflecting maximum congru-

ence with normal behavior, the other end reflecting maximum divergence, or pathological behavior. This distribution was normalized. Subjects for the four homogeneous groups were chosen randomly from patients having *T* scores between ± 1 standard deviation. The four heterogeneous groups each were comprised of two patients with *T* scores of less than -1 *SD*, two patients with *T* scores of greater than $+1$ *SD*, and three patients with *T* scores in the mid-range. Based on evidence presented by Bales and Borgatta (1955) and experience in group psychotherapy group size was limited to seven. Each of the four homogeneous and heterogeneous groups were assigned randomly to the treatment combinations of group psychotherapy and the drama group. All eight groups had the same average level of pathology as measured by the Lorr scale. There were no significant differences between groups regarding age, duration of hospitalization, or the taking of ataractics. Median age was 38.9 years. Median duration of hospitalization was 9.2 years. Fifty-three of the 56 experimental subjects were on ataractics.

Measures

The principal dependent variable, behavioral adjustment, was measured by the MACC Behavioral Adjustment Scale (Ellsworth, undated). The MACC produces scores entitled Motility, Affect, Cooperation, and Communication, and a Total Adjustment score, the sum of the Affect, Cooperation, and Communication scores. This scale has been shown to differentiate significantly between open and closed ward continued treatment patients, to be correlated significantly with the Hospital Adjustment scale and to be associated significantly with other measures of improved behavioral adjustment such as the length of time spent on pass. The scale is short, 14 items, and can be rated with high reliability. Pilot study data on ratings by pairs of raters used in the experiment produced interrater reliabilities ranging from .82 to .99 with an average coefficient of .92, taken over the 15 combinations of six rater pairs. These levels are consistent with reliabilities previously reported for the scale.

Ancillary measures of group cohesiveness and social choice were taken in the hope that the experimental groups would produce measurable changes in peripheral social behavior. The Semantic Differential profile given by each patient on himself was compared with the average profile he gave for other members of his group. A *D* statistic was calculated and interpreted as a measure of cohesion; a small *D* indicating cohesion.

Social choice data were obtained in a free choice situation. All patients on the experimental ward ate at the same time in the same area of the dining hall. They were seated four to a table but had complete freedom to choose any table in the area and any companions from among their fellow patients. Actual choice of companions at the noon meal was recorded for the 56 patients in the study. These choices were then categorized as "in group" or "out group" choices.

Method

Following the pilot study on reliability and the selection of groups, all subjects were rated on the MACC, were given the Semantic Differential (which included their name and the names of the other members of their group), and were observed at their noon meal for 3 consecutive days and their choice of companions recorded. Sleeping arrangements were changed so that group members had adjacent beds. Simultaneously the experimental procedures began. The four groups in group psychotherapy were seen twice each week for 1½ hours; a total of 3 hours a week. All groups were seen by the same therapist, the senior author.

The four drama groups were formed into two groups of 14, one homogeneous and one heterogeneous. In each group half of the patients were in group psychotherapy and half were not. These groups met three times a week for an hour. Generally they met in the Recreation Hall with a staff moderator, a recreational therapist from Special Services. This moderator had been instructed to supply the groups with all the material and information requested or needed by them for the production of plays, but to avoid taking over the "leadership" of the group. The role of the moderator might best be characterized as a "nondirective resource person." This role proved to be a difficult one to assume and was maintained only by frequent consultations between the experimenters and the moderator. Difficulties appeared to stem from the moderator's identification with the group himself. Thus, when a group once decided to put on a play reading from the scripts, the moderator became personally concerned over the adequacy of their decision. The moderator was present for all of the earlier meetings of the groups but missed some as time went on and occasional conflicting duties prevented his attendance. On those occasions the groups met without him. At the beginning of the study all patients in the drama groups were told individually that they had been chosen for a detail to provide plays for the entertainment of the staff and fellow patients. They were "assigned," not given a choice. Some patients protested leaving present details or simply engaging in an activity for which they did not care. Most patients, however, accepted the new "detail" with characteristic indifference. When complaints occurred about belonging to the group patients were told that although their dissatisfaction was understandable nothing could be done. Further, it was pointed out that they were obliged to meet this challenge but were free to do it in whatever way they decided as a group.

The study continued for 1 full year with measures taken every 6 weeks. The nurses and nursing assistants rating subjects on the MACC were not made aware of the specific hypotheses of the study or of the subgroups into which their patients fell. Two raters from different shifts were used for each rating period. Ratings were based on observations of the subjects for the week immediately preceding the date the ratings were due. It was agreed that any subject

leaving the hospital within 2 months after the beginning of the study was to be replaced by another randomly selected subject. Patients who left after longer than 2 months were counted as subjects but were replaced in groups by another "equivalent" patient. No data were collected on these "replacements" which were used only to maintain the groups at their full strength.

RESULTS

Because the primary dependent variable, the MACC, consisted of four subscores and a summary score, five separate analyses were done. In each case the analysis of the final scores was adjusted by multiple covariance for the effect of initial level and the length of time the subject was in the study. None of the main effects or interactions reached significance at the .05 level for the Affect subscores. The activity effect, however, with an F of 3.969, narrowly missed significance at this level. The activity effect was significant for all other subscores and for the total adjustment scores: motility, $p < .05$; communication, $p < .01$; cooperation, $p < .01$; and total adjustment, $p < .01$. Group psychotherapy reached significance at the .05 level for the motility subscores, but was nonsignificant for the other subscores and the total adjustment scores. The group structure effect did not reach significance on any of the measures. All interactions were nonsignificant.

Analysis of the Semantic Differential distance measures between self-rating and the average rating of other group members revealed no difference between original and final measures that could be attributed to any treatment or treatment combination. This measure produced very high attrition because of blank, incomplete, or obviously invalid protocols. It is interesting to note, however, that *all* distance measures used decreased over time and that this difference was significant at the .02 level.

Changes in choice of luncheon companions from outgroup to ingroup were practically nil. These social choices showed a remarkable consistency over time and no significant differences were obtained, either between treatment groups and combinations of treatment groups or between original and final choices over all subjects.

DISCUSSION AND CONCLUSIONS

The most compelling result of this study is the consistency with which the activity group showed significant change on the various categories of the MACC Behavioral Adjustment Scale. Changes were significant on all but the Affect subscale where the F missed significance at the .05 level by a value of only .08. These changes uniformly were in the direction of improved behavioral adjustment. The significant change in the Motility subscale reflected a lessening of motility. The data suggest this was a decrease in behavioral agitation and restlessness. Group psychotherapy showed a significant decrease in motility as well, but the data did not reach significance on the other subscales or on the Total Adjustment score. No significant results were attributable to the homogeneity-heterogeneity variable. During the study 18 patients left on trial visit or discharge, 2 of which returned within 6 months. No group or treatment showed a significant difference in this regard.

The fact that the significant differences found on the MACC for the activity group were not found in the Semantic Differential and social choice data for the same group reinforces earlier questions about these measures. A satisfactory method for screening invalid Semantic Differential protocols was not found. While some protocols were obviously invalid, e.g., those showing an invariant response pattern on the test form, in many cases this judgment was difficult to make. When the validity of a protocol remained in question it was accepted as data and treated as valid. This is an arbitrary procedure at best. Reliabilities on this instrument, using only the "valid" protocols, calculated from immediate test-retest by replicated items in the test form, ranged from $-.25$ to $.96$ with an average of $.58$. The overall change from pre- to posttest, if interpretable at all, most likely reflects a change in therapeutic procedures on the ward which occurred simultaneously but independently of the study. Overall rates of leaves, trial visits, and discharges also increased.

Although choice of luncheon companions was intended to be a measure of the formation of "real" groups resulting from the artificial experimental groupings, it became ob-

vious that this behavior was extremely stable and insensitive to change. Use of a behavioral measure which did not have a previously stereotyped pattern would have been advantageous.

The results of the study are encouraging. While only one treatment produced significant changes, it did so with compelling strength and consistency. The activity variable was responsible for most of the change in behavioral adjustment that occurred. It should be pointed out, however, that an important difference existed between the therapy and the drama groups in addition to the differences in "treatment." The group psychotherapist was a different person from the resource person associated with the drama groups. Thus, it is possible that the results document differences between the skills of these two people rather than between treatments as such. While this interpretation is possible it does not seem most parsimonious to the authors. This problem was evaluated when the study was designed and there appeared no feasible way of separating person effect from treatment effect and maintaining an adequate design. Additionally, the results favor the activity effect—a treatment wherein the resource person had only minimal contact. The amount and nature of contact with the subjects was specified as carefully as possible before the study began and every effort was made to insure they were maintained as such. Thus this problem does not affect the interpretation of the activity effect, which reached overall significance in any event. One could question the nonsignificance of the results in most areas for the group psychotherapy effect, however. It is possible that a more skilled therapist, using the same procedures, might have produced more significant results. It was decided to spell out the group psychotherapy procedures as clearly as possible and have all groups seen by the same therapist to avoid confounding intertherapist differences. Because the drama activity had its own discrete characteristics, in addition to the criteria specified in the study, it is impossible to state with certainty the source of the significant differences. It is clear, however, that the activity studied produced significant results in the predicted direction and there is reason to ex-

pect that it would do so again at another time or in another place.

This latter finding taken alone should be of significance to those in mental hospitals concerned with the treatment of chronic schizophrenics. The activity program studied here produced consistent and significant behavioral change with a minimum of staff intervention and expenditure of time. The "personnel efficiency" of such a treatment method is unquestionably of value. Of much greater significance, however, is the fact that this inexpensive technique produces results which, in this study, were incomparably better than the more "expensive" and time consuming group psychotherapy requiring a highly trained therapist. The implications of this study for the systematic use of nonprofessional personnel in the active treatment of chronic schizophrenics are compelling, as well as being attractive.

A number of refinements present themselves for future study. There is paramount need to vary activity programs by content, holding basic selection criteria constant, to evaluate the generality of the criteria. It would be expected, of course, that any activity program constructed to conform to the basic criteria and administered as the one currently studied would produce equivalent results. The difficult problem of therapist "effects" in group psychotherapy requires further attention. Although the homogeneity-heterogeneity variable did not produce significant results as it was studied, it is likely that the method of study could be improved. In this study it was advantageous to make the central tendency in the two types of group structure equivalent, varying only the dispersion. It is likely, however, that an effect due to group structure may interact with levels of pathology. Thus a study varying both effects systematically would provide informative data.

SUMMARY

Group psychotherapy, a specially designed activity program, and the homogeneity or heterogeneity of groups were evaluated as therapeutic modalities in the treatment of chronic schizophrenic patients. These variables were studied simultaneously in a $2 \times 2 \times 2$ factorial design with multiple covariance

of initial level of behavioral adjustment and length of stay in the therapeutic program. The primary dependent variable was the MACC Behavioral Adjustment Scale and its subscales. Measures of group cohesion and social choice also were obtained. The activity variable produced significant and consistent results in the predicted direction. Group psychotherapy produced relatively minor positive results and the group structure variable produced none. None of the interactions were significant. The ancillary measures of group cohesion and social choice showed no systematic change. The implications of this study for the use of this kind of activity program involving nonprofessional personnel in the treatment of chronic schizophrenic patients are positive and compelling. Refinements in design and suggestions for future research were presented.

REFERENCES

- BACH, G. R. *Intensive group psychotherapy*. New York: Ronald, 1954.
- BALES, R. F., & BORGATTA, E. F. Size of group as a factor in the interaction profile. In A. P. Hare, E. F. Borgatta, & R. F. Bales (Eds.), *Small groups*. New York: Knopp, 1955. Pp. 396-413.
- COHEN, L. B. The use of extramural activities in group psychotherapy with hospitalized female chronic schizophrenics. *Group Psychother.*, 1959, 12, 315-321.
- DIGIOVANNI, P. Orthodox group psychotherapy and activity group therapy with regressed schizophrenics. Unpublished doctoral dissertation, University of Illinois, 1958.
- ELLSWORTH, R. B. *MACC Behavioral Adjustment Scale*. Los Angeles: Western Psychological Services, undated.
- FRANK, J. D. Group psychotherapy with chronic hospitalized schizophrenics. In E. B. Brody & F. C. Redlich (Eds.), *Psychotherapy with schizophrenics*. New York: International Univer. Press, 1952. Pp. 216-230.
- GRAUER, D. Problems in psychotherapy with schizophrenics. *Amer. J. Psychother.*, 1955, 9, 216-233.
- HOFFMAN, L. R. Homogeneity of member personality and its effect on group problem-solving. *J. abnorm. soc. Psychol.*, 1959, 58, 27-32.
- KLAPMAN, J. W. *Group psychotherapy*. New York: Grune & Stratton, 1946.
- KLAPMAN, J. W. Didactic group psychotherapy with psychotic patients. In S. R. Slavson (Ed.), *The practice of group therapy*. New York: International Univer. Press, 1947.
- KRAMER, M. C. Group psychotherapy with psychotic patients. *J. nerv. ment. Dis.*, 1957, 125, 36-43.
- KRETCH, D., & CRUTCHFIELD, R. S. *Theory and problems of social psychology*. New York: McGraw-Hill, 1948.
- LAZELL, E. W. Group psychotherapy. In J. L. Moreno (Ed.), *Group psychotherapy: A symposium*. New York: Beacon, 1945.
- LORR, M. Multidimensional scale for rating psychiatric patients. *VA tech. Bull.*, 1953, No. 10-507.
- MURRAY, E. J., & COHEN, M. Mental illness, milieu therapy, and social organization in ward groups. *J. abnorm. soc. Psychol.*, 1959, 58, 48-54.
- MYERSON, A. Theory and principles of the "total push" method in the treatment of chronic schizophrenia. *Amer. J. Psychiat.*, 1939, 95, 1197-1204.
- OSGOOD, C. E., & SUCI, G. J. A measure of relation determined by mean difference and profile information. *Psychol. Bull.*, 1952, 49, 251-262.
- PAGE, R. E. Situational therapy. *J. Pers.*, 1957, 25, 578-588.
- PETERS, H. N. Learning as a treatment method in chronic schizophrenia. *Amer. J. occup. Ther.*, 1955, 9, 185-189.
- PEYMAN, D. A. R. An investigation of the effects of group psychotherapy on chronic schizophrenic patients. *Group Psychother.*, 1956, 9, 35-39.
- POWDERMAKER, FLORENCE B., & FRANK, J. D. *Group Psychotherapy*. Cambridge: Harvard Univer. Press, 1953.
- SCHER, J. M. Perception: Equivalence, avoidance, and intrusion in schizophrenia. *AMA Arch. Neurol. Psychiat.*, 1957, 77, 210-217. (a)
- SCHER, J. M. Schizophrenia and task orientation: The structured ward setting. *AMA Arch. Neurol. Psychiat.*, 1957, 78, 531-538. (b)
- SCHNADT, F. Techniques and goals in group psychotherapy with schizophrenics. *Int. J. group Psychother.*, 1955, 5, 185-193.

(Received August 24, 1960)

THE USE OF AN EXTENDED DRAW-A-PERSON TEST TO IDENTIFY HOMOSEXUAL AND EFFEMINATE MEN¹

LEIGHTON WHITAKER, JR.²

Wayne County General Hospital, Michigan

Classical psychoanalytic theory holds that individuals may have a psychological identification with the same sex or with the opposite sex (Fenichel, 1945). The concept of "psychosexual identity" has been used in projective testing of personality also. Regarding the Draw-A-Person Test (the DAP), Machover (1949, p. 101) has stated:

From the standpoint of sexual identification, it is assumed to be most normal to draw the self-sex first. From an empirical point of view, it is of interest that evidence of some degree of sexual inversion was contained in the records of all individuals who drew the opposite sex first in response to the instruction, "draw a person."

Presumably what is crucial in this instruction is that the individual is made to choose the sex of the person drawn and thereby projects his own psychosexual identity.

The present research represents a test of the theoretical and practical significance of choice of the sex of the figure in "free choice" drawings on the DAP by utilizing the choice as a psychometric sign to predict the characteristics "homosexuality" and "effeminacy" in men. As Meehl and Rosen (1955) have pointed out, the predictive efficiencies of such a psychometric sign must be evaluated relative to the base rates of the characteristics.

¹ Paper presented at the Michigan Academy of Arts, Sciences, and Letters, Psychology Division, Personality and Clinical Section, at Ann Arbor, Michigan, March 1960.

² Data for this research were collected while the author was at Recorder's Court Psychopathic Clinic, Detroit, with the kind permission of Alan Canty, Executive Director. Appreciation is expressed to John MacBride, formerly of the Court Clinic, and Bertram Cohen of Lafayette Clinic, Detroit, for their assistance in the beginning and end phases of the research, respectively.

METHOD

Two hundred and thirty-six men aged 16 to 65 with an average age of 28, who were referred to two clinical psychologists in a court clinic, served as subjects. Each subject was first given an examination which, at the minimum, consisted of a life-history interview and the Verbal section of the Wechsler Adult Intelligence Scale. Judgments were then made by the same examining psychologist, on the basis of the examination, as to whether the subject was homosexual or effeminate. Finally, the subject was given the extended DAP by the author. No attempt was made to select subjects and, with relatively few exceptions, all men referred to the two psychologists during a period of 7 months were used in the research.

Selection of the criteria according to which a subject was judged homosexual or effeminate presented three particular problems. First, the meaning of these terms varies considerably according to the individuals using them. It was necessary to give these terms highly explicit definitions so as to allow high criterion reliability. Second, the information on which the judgments were based varies somewhat in the kind and extent available. No special means was found to overcome this problem. Third, since not all homosexual men are effeminate according to psychoanalytic theory, separate judgments of homosexuality and effeminacy had to be made. However, it was assumed that most homosexual men do have a feminine psychosexual identity and it was expected that as a group, therefore, they would project this personality feature into their free choice drawings on the DAP.

The label "homosexual" meant admission by the subject to the examining psychologist of one or more instances of manifestly sexual impulses and/or behavior toward a person of the same sex when both parties were past pubertal age. The label "effeminate" meant one or more of the following: (a) display of effeminate speech, gestures, mannerisms, or dress during the examination; (b) personal description by the subject to the psychologist of playing a manifestly feminine role where, for 2 or more years, most of the subject's activity consisted of housework, sewing, washing, ironing, infant care, etc.; (c) personal description by the subject to the psychologist of very strong interests in typically feminine activities such

TABLE 1
DISCRIMINATION OF HOMOSEXUAL AND/OR EFFEMINATE MEN BY DRAW-A-PERSON SIGN

	Homosexual	Effeminate	H and/or E																																				
Regular DAP																																							
	Test - +	Test - +	Test - +																																				
Interview	<table> <tr> <td>+</td> <td>26</td> <td>19</td> <td>45</td> </tr> <tr> <td>-</td> <td>165</td> <td>26</td> <td>191</td> </tr> <tr> <td></td> <td>191</td> <td>45</td> <td>236</td> </tr> </table> $\chi^2 = 19, p < .001$	+	26	19	45	-	165	26	191		191	45	236	<table> <tr> <td>+</td> <td>16</td> <td>9</td> <td>25</td> </tr> <tr> <td>-</td> <td>175</td> <td>36</td> <td>211</td> </tr> <tr> <td></td> <td>191</td> <td>45</td> <td>236</td> </tr> </table> $\chi^2 = 5, p < .001$	+	16	9	25	-	175	36	211		191	45	236	<table> <tr> <td>+</td> <td>33</td> <td>26</td> <td>59</td> </tr> <tr> <td>-</td> <td>158</td> <td>19</td> <td>177</td> </tr> <tr> <td></td> <td>191</td> <td>45</td> <td>236</td> </tr> </table> $\chi^2 = 32, p < .001$	+	33	26	59	-	158	19	177		191	45	236
+	26	19	45																																				
-	165	26	191																																				
	191	45	236																																				
+	16	9	25																																				
-	175	36	211																																				
	191	45	236																																				
+	33	26	59																																				
-	158	19	177																																				
	191	45	236																																				
Extended DAP-A																																							
	Test - +	Test - +	Test - +																																				
Interview	<table> <tr> <td>+</td> <td>9</td> <td>36</td> <td>45</td> </tr> <tr> <td>-</td> <td>144</td> <td>47</td> <td>191</td> </tr> <tr> <td></td> <td>153</td> <td>83</td> <td>236</td> </tr> </table> $\chi^2 = 49, p < .001$	+	9	36	45	-	144	47	191		153	83	236	<table> <tr> <td>+</td> <td>4</td> <td>21</td> <td>25</td> </tr> <tr> <td>-</td> <td>149</td> <td>62</td> <td>211</td> </tr> <tr> <td></td> <td>153</td> <td>83</td> <td>236</td> </tr> </table> $\chi^2 = 27, p < .001$	+	4	21	25	-	149	62	211		153	83	236	<table> <tr> <td>+</td> <td>12</td> <td>47</td> <td>59</td> </tr> <tr> <td>-</td> <td>141</td> <td>36</td> <td>177</td> </tr> <tr> <td></td> <td>153</td> <td>83</td> <td>236</td> </tr> </table> $\chi^2 = 67, p < .001$	+	12	47	59	-	141	36	177		153	83	236
+	9	36	45																																				
-	144	47	191																																				
	153	83	236																																				
+	4	21	25																																				
-	149	62	211																																				
	153	83	236																																				
+	12	47	59																																				
-	141	36	177																																				
	153	83	236																																				
Extended DAP-B																																							
	Test - +	Test - +	Test - +																																				
Interview	<table> <tr> <td>+</td> <td>12</td> <td>33</td> <td>45</td> </tr> <tr> <td>-</td> <td>161</td> <td>30</td> <td>191</td> </tr> <tr> <td></td> <td>173</td> <td>63</td> <td>236</td> </tr> </table> $\chi^2 = 52, p < .001$	+	12	33	45	-	161	30	191		173	63	236	<table> <tr> <td>+</td> <td>7</td> <td>18</td> <td>25</td> </tr> <tr> <td>-</td> <td>166</td> <td>45</td> <td>211</td> </tr> <tr> <td></td> <td>173</td> <td>63</td> <td>236</td> </tr> </table> $\chi^2 = 30, p < .001$	+	7	18	25	-	166	45	211		173	63	236	<table> <tr> <td>+</td> <td>17</td> <td>42</td> <td>59</td> </tr> <tr> <td>-</td> <td>156</td> <td>21</td> <td>177</td> </tr> <tr> <td></td> <td>173</td> <td>63</td> <td>236</td> </tr> </table> $\chi^2 = 80, p < .001$	+	17	42	59	-	156	21	177		173	63	236
+	12	33	45																																				
-	161	30	191																																				
	173	63	236																																				
+	7	18	25																																				
-	166	45	211																																				
	173	63	236																																				
+	17	42	59																																				
-	156	21	177																																				
	173	63	236																																				
Extended DAP-C																																							
	Test - +	Test - +	Test - +																																				
Interview	<table> <tr> <td>+</td> <td>29</td> <td>16</td> <td>45</td> </tr> <tr> <td>-</td> <td>182</td> <td>9</td> <td>191</td> </tr> <tr> <td></td> <td>211</td> <td>25</td> <td>236</td> </tr> </table> $\chi^2 = 37, p < .001$	+	29	16	45	-	182	9	191		211	25	236	<table> <tr> <td>+</td> <td>18</td> <td>8</td> <td>25</td> </tr> <tr> <td>-</td> <td>193</td> <td>17</td> <td>211</td> </tr> <tr> <td></td> <td>211</td> <td>25</td> <td>236</td> </tr> </table> $\chi^2 = 13, p < .001$	+	18	8	25	-	193	17	211		211	25	236	<table> <tr> <td>+</td> <td>38</td> <td>21</td> <td>59</td> </tr> <tr> <td>-</td> <td>173</td> <td>4</td> <td>177</td> </tr> <tr> <td></td> <td>211</td> <td>25</td> <td>236</td> </tr> </table> $\chi^2 = 52, p < .001$	+	38	21	59	-	173	4	177		211	25	236
+	29	16	45																																				
-	182	9	191																																				
	211	25	236																																				
+	18	8	25																																				
-	193	17	211																																				
	211	25	236																																				
+	38	21	59																																				
-	173	4	177																																				
	211	25	236																																				

as the above. A subject was rated homosexual or effeminate only if the evidence was unequivocal.

The two psychologists who examined and rated the subjects each had at least 3 years training and/or experience beyond the master's degree in clinical psychology. To provide an estimate of the reliability of their ratings, each psychologist independently rated the first 40 subjects for homosexuality and effeminacy. There was complete agreement in rating homosexuality and agreement in 38 of the 40 cases in rating effeminacy.

The extended version of the DAP was administered by the author as follows: Each man was told "draw a person." When the sex of the figure drawn was decided by the man, he was told "Now draw a person of the opposite sex." Finally, a third drawing was obtained where, for the second time, the man himself chose the sex of the figure. The test procedure thus extended the regular DAP where the subject produces only two drawings with the sex of only one of these chosen by himself. It was possible, therefore, to compare the discriminating powers of the regular and extended versions of the test.

Since each man produced two free choice drawings, i.e., drawings which could be either of male or female figures, there were four possible ways in which a man's DAP might be scored positive in psychometric sign. In the regular version of the test a positive psychometric sign is scored when the first drawing is of a female figure. There are three additional methods of scoring possible with the extended version of the test. Method A scores a test protocol positive in psychometric sign if either free choice drawing is female. Method B scores positive if the second free choice is female. Method C scores positive if both free choices are female.

RESULTS AND DISCUSSION

As shown in Table 1, all four methods of scoring psychometric sign discriminated the characteristics homosexual, effeminate, and homosexual and/or effeminate beyond chance

levels. This aspect of the results supports the theoretical expectation, based upon psychoanalytic and projective test concepts of psychosexual identity, that psychosexual identity is projected into the choice of sex of the figures drawn in free choice drawings on the DAP.

As shown in Table 2, when the efficiencies of the various signs are compared with the efficiencies of classifying everyone simply as *not* possessing the characteristic in question, it is clear that the signs have no appreciable value except for predicting the characteristic homosexual and/or effeminate. Even here the improvement is most modest, at best 84% vs. 75%. However, differential weighting of false positives vs. false negatives might alter conclusions, dependent upon which error was judged worse. For example, if it were of primary importance to screen out all men with the characteristics and of only secondary importance to avoid screening out some men without the characteristics, then the psychometric signs may be of practical use. As shown in Table 3, the extended DAP with Scoring Method A screened 80%, 84%, and 80% of the men with the characteristics homosexual, effeminate, and homosexual and/or effeminate, respectively. At the same time 57%, 75%, and 43% of the men without these respective characteristics were screened out.

In view of the potential usefulness of the extended DAP as a screening device, it is suggested that the test be tried in other settings where cross-validation data could be ob-

TABLE 2
PREDICTIVE EFFICIENCIES OF THE PSYCHOMETRIC SIGNS AND BASE RATES IN TERMS OF
CORRECT DECISIONS FOR INDIVIDUALS WITHIN THE GROUPS
(N = 236)

Correct decisions	Homosexual		Effeminate		H and/or E	
	N	%	N	%	N	%
Base rates ^a	191	81				
Regular DAP	184	78	211	89		
Extended DAP-A	180	76	184	78	177	75
Extended DAP B	194	82	170	72	191	81
Extended DAP C	198	84	184	78	188	80
			201	85	198	84
					195	93

Note. Base frequency for Homosexuals = 45; Effeminate = 25; H and/or E = 59.
^a N and percentage of correct decisions if all subjects are classified in the more probable category, i.e., *not* homosexual and/or *not* effeminate.

TABLE 3
MEN WITH THE CHARACTERISTICS WHO ARE POSITIVE IN PSYCHOMETRIC SIGN
($N = 236$)

Test	Homosexual		Effeminate		H and/or E	
	<i>N</i>	%	<i>N</i>	%	<i>N</i>	%
Regular DAP	20	44	8	32	27	46
Extended DAP-A	36	80	21	84	47	80
Extended DAP-B	33	73	18	72	42	71
Extended DAP-C	16	36	7	28	21	36

tained. Perhaps the greatest obstacle to obtaining such data will be the difficulty, characteristic of attempts to establish the validity of projective techniques, in finding adequate criterion measures. Some observations in the present research suggest that more adequate criterion measures would have shown the extended DAP to be a more powerful discriminator of psychosexual identity than the tables of results indicate. For example, of the nine homosexual men who were not positive on the extended DAP with Scoring Method A five limited their homosexual activity to playing the "masculine" role and would not, therefore, be expected to have a feminine psychosexual identity according to psychoanalytic theory (Fenichel, 1945).

SUMMARY

Two hundred and thirty-six men, referred to a court clinic, were rated on the characteristics homosexuality and effeminacy by a clinical psychologist on the basis of life-history interviews which he conducted. Each man was then given an extended Draw-A-Person Test on which he chose the sex of two of the three figures drawn. All four possible methods of scoring a test protocol as "positive" in

psychometric sign (one or more free choice drawings of a female) were used to predict the characteristics. The results support the theoretical expectation, based on psychoanalytic and projective test concepts of psychosexual identity, that psychosexual identity is projected into free choice drawings. The psychometric signs were not more efficient, overall, than the base rates in predicting the characteristics. However, differential weighting of false positives vs. false negatives might alter conclusions about the practical usefulness of the signs, dependent upon which error was judged worse. It was suggested that the extended Draw-A-Person Test be used in other settings where cross-validation data could be obtained.

REFERENCES

- FENICHEL, O. *The psychoanalytic theory of neuroses*. New York: Norton, 1945.
- MACHOVER, KAREN. *Personality projection in the drawing of the human figure*. Springfield, Ill.: Charles C Thomas, 1949.
- MEEHL, P. E., & ROSEN, A. Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychol. Bull.*, 1955, 52, 194-216.

(Received August 31, 1960)

THE PSYCHOLOGICAL SIGNIFICANCE OF THE MMPI K SCALE IN A NORMAL POPULATION

ALFRED B. HEILBRUN, JR.

State University of Iowa

The long-standing assumption that the *K* scale is a measure of defensiveness stems directly from the nature of its derivation—the detection of hospitalized psychiatric patients who presented normal profiles on the MMPI. However, two lines of evidence have suggested that *K* scale scores may also relate to general level of adjustment. For one, Wheeler, Little, and Lehner (1951) found that normal groups scored higher on the *K* scale than abnormal groups, and their interpretation of these findings was consistent with general clinical practice which has for a number of years stressed *K* scale elevations as an index of ego adequacy. The other line of evidence is the rather consistent finding that *K* scale scores show an increase when posttreatment MMPI scores are compared with pretreatment (Carp, 1950; Feldman, 1952; Gallagher, 1953; Hales & Simon, 1948; Schofield, 1953).

More recently, Smith (1959) has provided evidence which suggests the *K* scale is not an adequate measure of defensiveness for normal populations. He found a significant negative correlation ($-.39$) between *K* scores and judgments of defensiveness made by subjects with 40 hours of observation on which to base their ratings. Smith used the results of his study to argue that "it is defensive for abnormal population Ss to obtain high *K* scale scores but a sign of health for normal population Ss" (p. 276).

If true, the implications of a differential psychological meaning for *K* scale scores for abnormal and normal populations are serious. The original addition of a *K* increment to five of the nine MMPI clinical scales followed a demonstrated enhancement of discrimination between largely inpatient psychoneurotic and psychotic groups and a general Minnesota "normal" group. However, extension of this

weighting system to personality measurement within a normal population assumes that (a) *K* measures defensiveness, (b) defensiveness is associated with a lowering of MMPI scale scores, and (c) the correction by adding a *K* increment raises these scale scores and provides a more veridical assessment of the person. However, if *K* is positively correlated with psychological adjustment for normal subjects and is not a measure of defensiveness, the *K* correction would appear to be operating in direct opposition to test validity. That is, higher *K* scores would tend to be associated with better adjustment for normal subjects; yet the higher the *K* scale score, the greater the *K* correction and the more the elevation of the clinical scales in the psychopathological direction. The problem of appropriate *K* usage in normal population testing was foreseen by the original workers (McKinley, Hathaway, & Meehl, 1948) with the MMPI who stated:

For other clinical purposes it is possible that other λ -values [i.e., *K* weights] would be more appropriate. Thus, it seems likely that for the best separation of "maladjusted normals," such as those which abound in a college counseling bureau . . . , other weights might be better (p. 24).

Results such as those provided by Smith's study which raise questions about the appropriateness of standard MMPI usage in normal populations warrant careful scrutiny. It is the purpose of the present study to evaluate two hypotheses suggested by Smith. These are:

1. The *K* scale is a measure of psychological health (adjustment) in a normal population.
2. The *K* scale is not a measure of defensiveness in a normal population.

METHOD

Definition of Differing Adjustment Groups within a Normal Population

To test the hypothesis that *K* scale performance is positively related to level of psychological health within a normal population, two types of groups from a normal college population clearly differing in adjustment level were selected. A group of college males ($N=146$) who had sought help at the University Counseling Service, State University of Iowa, constituted a poorer adjusted normal (PA) group. This number included 100 subjects who had requested vocational and/or educational counseling and 46 subjects who sought counseling for personal adjustment problems. The male better adjusted (BA) group was comprised of 153 students, none of whom had been seen in the Counseling Service.

Parallel female PA and BA groups were also constituted. The female PA group included 143 Counseling Service clients, 100 having requested help for vocational and/or educational problems and the remaining 43 for personal adjustment problems. The female BA subjects ($N=197$) were students who had not requested help at the Counseling Service. Both the male and female PA groups approximate representative samples of Counseling Service clients as far as the proportions of vocational-educational and personal adjustment counseling requests are concerned.

Each of the 639 subjects included in the male and female BA and PA groups had taken the MMPI under one of two conditions: (a) as part of the university prefreshman entrance battery in the summer of 1958 or 1959, or (b) as part of the Counseling Service intake battery. All of the BA subjects took the MMPI as part of the prefreshman battery, and about 90% of the subjects in the PA groups did. Thus, the *K* scale scores for the four adjustment groups are provided by very homogeneous samples relative to both age and educational level at time of testing. Since there is some evidence (Sarason, 1956) that *K* scores are related to intellectual ability in college subjects, separate preliminary analyses of this relationship for the males and females in the present samples were conducted. The product-moment correlation between *K* and the mean composite percentile on the university entrance examination for both males ($N=319$) and for females ($N=373$) was .11, significant at the .05 level. Although this figure suggests a relationship of limited magnitude, the four adjustment groups were matched for ability level. The group composite percentile means were: male PA = 53.91, male BA = 53.72, female PA = 54.47, female BA = 54.60.

Based on Smith's hypothesis that the *K* scale measures psychological health in a normal population, it was predicted that the BA male and female groups would score higher on the *K* scale than would the PA groups, level of adjustment being defined in terms of soliciting or not soliciting help for psychological problems subsequent to testing.

Definition of Defensiveness

For purposes of his investigation, Smith used the definition of defensiveness provided by Page and Markowitz (1955) as follows: "The defensive individual is described as one who fails to ascribe to himself characteristics of a generally valid but socially unacceptable nature" (p. 431). In the present study the concept of defensiveness was extended so as to include the self-ascription of characteristics which are not valid but are socially acceptable as well as the denial of valid but unacceptable characteristics.

Since adjustive behaviors are much more likely to be socially acceptable than nonadjustive behaviors, one difficult aspect of deriving a useful measure of defensiveness in a normal population is the increased probability that a socially acceptable self-description is also factually correct. In the present study this problem of confounding defensiveness and accurate self-description was approached by using the self-descriptions of a group of subjects at the maladjusted end of a normal population adjustment range. Self-descriptions on the 300-item Adjective Check List (ACL) (Gough, 1960) of 50 male college students who sought help for personal adjustment problems at the Counseling Service were scored for the number of endorsed adjectives which were included in the 75 judged to reflect most favorably on the endorser and the number included in the 75 judged to reflect least favorably on the endorser (see Gough, 1955). Subtracting the latter count from the former provided a "favorability" count for each subject. By cutting the distribution of favorability scores at the median, a group of subjects giving more favorable self-descriptions and a group giving less favorable self-descriptions were obtained. Since all subjects were maladjusted it was assumed that the subjects giving more favorable self-descriptions represented a more defensive group. A "Defensiveness" (Def) scale was then developed for the ACL by determining through chi square analysis which adjectives out of the 300 total reliably discriminated between the high and low favorability subjects. The 61 adjectives which discriminated were cross-validated on a new sample of 34 male personal adjustment counseling subjects, and 28 of these adjectives significantly differentiated the newly constituted high and low favorability groups. These 28 adjectives¹ (composed of 27 favorable adjectives and one unfavorable adjective which became a subtractive item) were included in the male Def scale. It can be noted that the mode of derivation for the male Def scale (as well as the female scale described below) parallels that of the *K* scale in the inclusion of items which discriminated between maladjusted subjects who portrayed themselves psychometrically in

¹ The lists of adjectives included in the male and female Defensiveness scales may be obtained without charge from Alfred B. Heilbrun, Jr.; Department of Psychology, State University of Iowa; Iowa City, Iowa.

TABLE 1

MEANS AND *SD*s OF MALE AND FEMALE CORRECTED DEFENSIVENESS SCALES SCORED ON ACLs OBTAINED UNDER STANDARD, DEFENSIVE, AND IDEAL-SELF CONDITIONS

Testing condition	<i>N</i>	Male mean	<i>SD</i>	<i>N</i>	Female mean	<i>SD</i>
Standard	97	15.30	4.58	114	16.85	5.34
Defensive	30	18.63	3.92	11	22.18	4.23
Ideal-self	56	21.00	.79	25	24.32	2.15

an unduly favorable light and those subjects who did not.

It was found that for a sample of 97 normal male students the Def score was positively correlated with total number of adjectives checked ($r = .71$), so a correction was necessary. The mean number of endorsed adjectives for these subjects was 95.16 with an $SD = 30.08$, while the SD for the Def scale was 3.28. A correction of one point (about one-third SD) on the subjects Def scale score was made for each 10 endorsed adjective (about one-third SD) deviations of total checked from the group mean of 95—added if the total number was below this figure and subtracted if the total fell above 95. Six to nine adjectives were counted as a point. This correction successfully removed the contingency of Def scores upon total adjective endorsement as attested by the insignificant correlation of $r = .08$ between these variables for a new group of male college students ($N = 109$). The reliability of the male corrected Def scale scores is .67, based on a 10 week test-retest sample of 43 college subjects.

A defensiveness measure for the ACL was derived for female college subjects by the same procedures as for the males. The original group of maladjusted Counseling Service subjects consisted of 43 females while the replication sample included 55 subjects. Seventy-two adjectives reliably discriminated between high and low favorability groups in the original sample, while 36 items held up on cross-validation and were included in the female Def scale. Of this number, 28 were favorable adjectives and 8 were unfavorable, thus being subtracted from the cumulative score. There were 10 overlapping items between the male and female scales.

A similar positive correlation ($r = .66$) between female Def scale scores and number of adjectives checked was found for a sample of 114 college females. Correction was made based on the following statistics: mean number of adjectives checked = 92.63; SD of total adjectives checked = 30.75; and SD of Def scale = 6.97. It was found that using the adding or subtracting of two Def scale points (about one-third SD) for each 10 adjectives checked (about one-third SD) above or below 93 adjectives (with partial credits for parts of 10) overcorrected and produced a negative correlation of $-.64$ between Def scores and total adjectives. The same correction used with the male scale was then applied and quite

successfully eliminated the relationship ($r = -.03$) between Def scores and total adjectives checked, based on the performances of the 103 additional female college subjects. The 10 week test-retest reliability of the Def scale for females ($N = 56$) is .79.

Some preliminary evidence was available to evaluate the male and female Def scales as measures of defensiveness. These scales were scored on ACLs administered to normal college subjects under three conditions: (a) a standard instruction research condition; (b) a standard instruction defensiveness-inducing condition (Heilbrun, 1958); and (c) an ideal-self-description instruction condition (Heilbrun, 1958). If Def scales are measures of defensiveness, scores might be expected to show a progressive increase over these three conditions. Table 1 shows this expected progressive mean increase as well as an increasing homogeneity of scores for Def scores. Tests of significance showed the difference in standard and defensive condition means for males ($t = 3.15$ for 123 df ; $p < .01$) and females ($t = 3.80$ for 127 df ; $p < .001$) to be highly significant. Various considerations (e.g., overlapping subjects, heterogeneous variances) made testing for mean differences between defensive and ideal-self conditions unfeasible. Also, to evaluate whether the Def scales may be reflecting differences in adjustment level rather than defensiveness, the Def scale means for groups of more poorly adjusted Counseling Service males (15.26; $SD = 5.25$; $N = 109$) and females (16.25; $SD = 6.05$; $N = 103$) were compared to the scale means of the normal (i.e., "standard condition") male (15.30) and female (16.85) groups reported in Table 1. The almost identical means indicate that Def scores are not measures of adjustment level.

Despite the preliminary evidence that the Def scales do measure defensiveness in self-description for normal college population subjects, it is clear that scores on these scales still confound defensiveness and true adjustment level (i.e., some high scorers are truly well adjusted and some low scorers are truly maladjusted). The most that is contended is that proportionally more of the performance variance on the Def scales can be attributed to defensiveness than would be the case for performance on the entire ACL, the proportions in either case being unspecified.

To test Smith's hypothesis that the K scale is not a measure of defensiveness in a normal population,

two correlational analyses were conducted: (a) the K scores of 103 Counseling Service female and 109 Counseling Service male college students were correlated with their Def Scores on the ACL, and (b) the K scores of 141 normal college female and 92 normal college male subjects were correlated with their ACL Def scores. ACLs for all subjects in the Counseling Service groups were administered as part of the Counseling Service intake battery at a variable time (from a few moments to almost 2 years) following the administration of the MMPI. The normal subjects were given the ACL under research conditions from 1 to 18 months after taking the MMPI.

RESULTS AND DISCUSSION

Hypothesis I: The K scale is a measure of psychological health in a normal population

The mean K standard score for the male PA group was 55.62 ($SD = 8.55$) compared to a mean K standard score of 54.18 ($SD = 8.35$) for the BA male subjects. These mean values do not differ significantly from each other ($t = 1.43$ for 299 df ; $.15 < p < .20$). The female PA group had a mean K score of 56.83 ($SD = 7.25$) whereas female BA subjects had a mean K score of 58.02 ($SD = 7.02$). Again the difference in means did not differ reliably from zero ($t = 1.51$ for 338 df ; $.10 < p < .15$). Thus, there is no support in these data for the contention that the K scale measures degree of psychological adjustment in a normal population. However, since each of the PA groups was composed of two subsets of subjects differing in level of adjustment (i.e., better adjusted vocational-educational cases vs. more poorly adjusted personal adjustment cases), it remained a possibility that differences in K might be demonstrated if more extreme comparisons were made. Accordingly, the mean K scores for the personal adjustment counseling subjects and the BA subjects were compared. For males, the mean K score for the personal adjustment subjects ($N = 46$) was 54.24 ($SD = 8.62$), whereas this mean score for the BA subjects ($N = 153$) was an almost identical 54.18 ($SD = 8.35$). For females, the mean K score for the personal adjustment subjects ($N = 43$) was 55.23 ($SD = 6.35$) and the K scale mean for the BA subjects ($N = 197$) was 58.02 ($SD = 7.02$). This difference was significant at the .05 level of confidence ($t = 2.36$ for 238 df). Thus, there is some evi-

dence that K is positively related to level of psychological adjustment for females when extreme groups are compared but no evidence for this relationship with males.

Since all PA subjects were Counseling Service clients who had solicited help, one possible bias in these K by adjustment level analyses is that such "help-seekers" would also tend to be uncritically frank endorsers of pathology (i.e., "plus getters") on the MMPI. Since plus-getting should be associated with lowered K scores, such a bias would operate in the direction of supporting the hypothesis that the more poorly adjusted subjects in this study would have lower K scores than normal subjects who had not sought psychological help. There does not appear to be any way to analyze the possible effect of plus getting within the current data, although it can be noted that Hypothesis I failed to receive any support in the male analysis despite this possible bias effect. It might be added that the lowered circumspection implicit in plus getting behavior can actually be considered a part of the true pathology of subjects, representing as it does a marked lowering of the ego defenses.

Hypothesis II: The K scale is not a measure of defensiveness in a normal population

The product-moment correlation between K standard scores and scores on the Def scale was .43 for the 109 male Counseling Service subjects. Considering the only moderate test-retest reliabilities of the two scales² and the fact that considerable time typically elapsed between the two tests, a correction for attenuation was applied and this correlation became .64. Both correlations are significant beyond the .01 level of confidence. For the 103 Counseling Service female subjects, the correlation between K and Def was .26. After correction for attenuation this correlation was .35, both figures being significant beyond the .01 level of confidence. The correlation between K and Def scale scores for normal college males ($N = 92$) was .24 ($p < .05$) or .35 ($p < .01$) corrected for attenuation. This cor-

² A test-retest reliability figure of .70 was used for the K scale in all corrections for attenuation. This figure has been suggested as a best estimate (Dahlstrom & Welsh, 1960, p. 53).

relation was $-.25$ ($p < .01$) for normal college females ($N = 141$) or $-.36$ ($p < .01$) following correction.

These sets of correlations suggest two things. For one, the K scale appears to be a better measure of defensiveness for maladjusted subjects in a normal population than for the better adjusted subjects in such a population. The decrease in the positive relationship between the K scale and the measures of defensiveness comparing the correlation for the male maladjusted group to that for the male adjusted group (.29) and the correlation for the female maladjusted group to that for the female adjusted group (.71) were both significant ($p < .05$ and .001, respectively). This finding is generally consistent with Smith's (1959) argument that "it is defensive for abnormal population Ss to obtain high K scale scores but a sign of health for normal population Ss" (p. 276), if the reasoning is extended to maladjusted vs. adjusted subjects with a normal population. The second implication of these correlational data is that a sex difference in the psychological meaning of K scale performance may exist. In the analyses of maladjusted normal groups, the females provided a significantly positive correlation between K scale performance and a measure of defensiveness but one that was reliably lower ($p < .05$) than that for the males. When adjusted normal groups were analyzed, males continued to show a significant positive relationship between their K scale scores and Def scale scores, whereas females showed the opposite pattern—the higher the K scale score tended to be, the *lower* the defensiveness. The reversal for adjusted college females of the usual psychological significance attributed to the K scale was also found by Smith in his predominantly male group of industrial supervisors. This reversal, taken in conjunction with the finding that adjusted females showed significantly higher K scores than more seriously maladjusted females, suggests that both of Smith's hypotheses received considerably more support from the female data than from the male.

In conclusion, the data from the present study indicate that the K scale is positively related to defensiveness when more maladjusted subjects from a normal college popula-

tion are appraised but is less positively related or even negatively related to defensiveness when psychologically healthy subjects are considered. The alternative psychological implication of K for normal population subjects suggested by Smith—degree of psychological health—received some support in the case of females but none in the case of males. These results suggest that the standard K correction for the MMPI clinical scales is psychometrically advantageous in normal college population testing with male maladjusted subjects, somewhat less useful with maladjusted females and better adjusted males, and a source of invalidity with better adjusted females.

SUMMARY

Two hypotheses taken directly from a study by Smith (1959) and indirectly from earlier investigators were tested in the present study: (a) the K scale of the MMPI is a measure of psychological health in a normal population, and (b) the K scale is not a measure of defensiveness in a normal population. To test Hypothesis I, the K scores of two samples of maladjusted male ($N = 146$) and female ($N = 143$) Counseling Service clients were compared with the K scores of male ($N = 153$) and female ($N = 197$) college normals. No significant differences were found in either comparison, although the normal female group mean K score was reliably higher than that of a subgroup of the most seriously maladjusted females ($N = 43$). Thus, there was some support for the hypothesis that K is a measure of psychological health in a normal population in the case of females only.

Hypothesis II was tested by correlating K scale scores with specially constructed male and female Defensiveness scales for the Adjective Check List. Significant correlations of .64 for male Counseling Service subjects ($N = 109$) and .35 for female Counseling Service subjects ($N = 103$) support the assumption that the K scale is a measure of defensiveness for maladjusted subjects in a normal population. However, when these relationships were determined for normal male ($N = 92$) and female ($N = 141$) college subjects, reliably different correlations between K and the de-

defensiveness measures were obtained (.35 and -.36, respectively). These correlational data tended to support Smith's contention that the K scale is a better measure of defensiveness among more maladjusted subjects.

REFERENCES

- CARP, A. MMPI performance and insulin shock therapy. *J. abnorm. soc. Psychol.*, 1950, 45, 721-726.
- DAHLSTROM, W. G., & WELSH, G. S. *An MMPI handbook: A guide to use in clinical practice and research*. Minneapolis: Univer. Minnesota Press, 1960.
- FELDMAN, M. J. The use of the MMPI profile for prognosis and evaluation of shock therapy. *J. consult. Psychol.*, 1952, 16, 376-382.
- GALLAGHER, J. J. MMPI changes concomitant with client centered therapy. *J. consult. Psychol.*, 1953, 17, 334-338.
- GOUGH, H. G. Reference handbook for the Gough Adjective Check List. Berkeley: University of California, Institute for Personality Assessment & Research, 1955. (Mimeo)
- GOUGH, H. G. The Adjective Check List as a personality assessment research technique. *Psychol. Rep.*, 1960, 6, 107-122.
- HALES, W. M., & SIMON, W. MMPI patterns before and after insulin shock therapy. *Amer. J. Psychiat.*, 1948, 105, 254-258.
- HEILBRUN, A. B., JR. Relationships between the adjective check list, personal preference schedule, and desirability factors under varying defensiveness conditions. *J. clin. Psychol.*, 1958, 14, 283-287.
- McKENLEY, J. C., HATHAWAY, S. R., & MEEHL, P. E. The MMPI: VI. The K scale. *J. consult. Psychol.*, 1948, 12, 20-31.
- PAGE, H. A., & MARKOWITZ, GLORIA. The relation of defensiveness to rating scale bias. *J. Psychol.*, 1955, 40, 431-435.
- SARASON, I. G. The relationship of anxiety and "lack of defensiveness" to intellectual performance. *J. consult. Psychol.*, 1956, 20, 220-222.
- SCHROFIELD, W. A. A further study of the effects of therapies on MMPI responses. *J. abnorm. soc. Psychol.*, 1953, 48, 67-77.
- SMITH, E. E. Defensiveness, insight, and the K scale. *J. consult. Psychol.*, 1959, 23, 275-277.
- WHEELER, W. M., LITTLE, K. B., & LEHNER, G. F. J. The internal structure of the MMPI. *J. consult. Psychol.*, 1951, 15, 134-141.

(Received September 8, 1960)

SELF-SATISFACTION AND PSYCHOLOGICAL ADJUSTMENT IN SCHIZOPHRENICS

DENNIS K. KAMANO¹

Galesburg State Research Hospital, Illinois

It is generally recognized that the satisfaction or concern of an individual with his phenomenal self represents an important aspect of psychological adjustment. For example, it has been demonstrated that marked dissatisfaction with one's self is indicative of conflicts or maladjustment (Cowen, Heiliger, & Axelrod, 1955), while positive and self-accepting attitudes towards the self are associated with good psychological adjustment (McQuitty, 1950; Rogers, 1950). Most of these studies, however, were confined primarily to normal and psychoneurotic subjects, but the relationship between self-satisfaction and psychological adjustment in regard to other classes of individuals is not clear. It follows that a particular formulation found to be applicable to one class of subjects may not be applicable, in the same way, to other classes of individuals. Such paradox in conception may be seen in the widely accepted view that manifest anxiety is patently disruptive and maladaptive when seen in a nonhospitalized individual but is a prognostically good sign when seen in a hospitalized patient (Arieti, 1955). Similarly, it is granted that to admit satisfaction with one's self is indicative of good adjustment in a normal individual, but is it a prognostically good sign when seen in hospitalized schizophrenic patients? The relationship between self-satisfaction and psychological adjustment is a complex one, and there is a need for further study and a qualified interpretation.

It has been widely recognized by clinicians that schizophrenic patients differ markedly in their degree of expressed self-satisfaction and adaptive potential. In contrast to normal subjects, it has been widely recognized by cli-

nicians that with hospitalized schizophrenic patients at least, concern with one's self represents a more adaptive behavior than self-satisfaction when this is based upon suppressive and repressive mechanisms. Some patients reveal extreme self-satisfaction, a frequently observed behavior used by patients to deny to themselves the extent of their discontent and pathology. Such patients are likely to be unrealistic, inflexible, and resistant towards any forces threatening to disrupt such rigid self-definition to such an extent that adaptive potential is grossly reduced. Much depends, of course, on the concept of adjustment one subscribes to. Most psychologists would agree in considering a suppressive, repressive mode of adaptation in hospitalized schizophrenics as less than adequate. Such behavior may represent a condition sufficient enough for a stable and benign hospital environment where pressure on the patient never becomes too great, but one which is incapable of manifesting adaptive flexibility in other situations. In a sense, such a person is adapted, as far as hospital adjustment is concerned, but not adaptable.

The above considerations led to the formulation of the following hypotheses with which the present study is principally concerned:

1. Schizophrenic subjects revealing extreme self-satisfaction will tend to deny and suppress threatening features of themselves to such an extent that this will be reflected in their response to a personality evaluation. That is, schizophrenic subjects revealing extreme self-satisfaction will reveal lower recall of unfavorable personality characteristics from a passage designed to simulate a personality evaluation, as compared with schizophrenic subjects not so characterized. Implicit in this proposition is the corollary that extremely self-satisfied schizophrenic subjects

¹ The author wishes to express his appreciation to Janet E. Drew and Vasso Vassiliou for their assistance in the collection of the data.

will tend to reveal higher recall of favorable items consistent with their highly favorable self-concept.

2. Schizophrenic subjects revealing extreme self-satisfaction based upon denial and repressive mechanisms, represent a state of unrealistic self-appraisal and general reduction in their capacity for evaluation, and such reduction in evaluative ability will be reflected in a situation requiring realistic appraisal of their performance from an objective frame of reference. That is, schizophrenic subjects revealing extreme self-satisfaction will reveal greater discrepancy between their level of performance and level of aspiration than schizophrenic subjects not so characterized. Schizophrenic subjects admitting some dissatisfaction with themselves will tend to set their level of aspiration more in relation to their actual level of performance and reveal lower discrepancy scores than extremely self-satisfied schizophrenic subjects.

METHOD

Measure of Self-Satisfaction

There are several ways in which self-regarding attitudes can be measured. One method is to use the semantic differential technique and to index the evaluation of the self-concept along the scale provided by the subject's own judgments of the concepts, my Actual Self (AS), my Ideal-Self (IS), and my Least-Liked Self (LLS), e.g., the distance from AS to LLS as a ratio to the total distance from LLS to IS (Osgood, Suci, & Tannenbaum, 1957). This ratio, LLS-AS/LLS-IS, was used in this study as an index of self-satisfaction. The ratio, LLS-AS/LLS-IS, approaches 1.00 as the location of AS approaches that of IS, i.e., as one's self-satisfaction increases. In other words, the value increases in size with self-satisfaction.

These self-concepts were rated on 15 bipolar scales which were presumed to be relevant. The scales used included 6 representative of the evaluative factor (attracting-repelling, complete-incomplete, important-unimportant, healthy-sick, high-low, and sociable-unsociable), 5 for the potency factor (large-small, hard-soft, strong-weak, deep-shallow, and masculine-feminine), and 4 for the activity factor (active-passive, hot-cold, tense-relaxed, and aggressive-defensive).

Subjects

Forty-four institutionalized white women labeled as schizophrenics were subjects in the present study. The subjects were screened to determine that they were all sufficiently in contact and of adequate op-

erating intelligence to understand and complete the tasks involved. Self-satisfaction scores for the 44 subjects were secured from the ratio, LLS-AS/LLS-IS, by averaging the ratio for each of the 15 scales. There were 17 subjects who received a score of 1.00 or more, while 27 subjects received a score of less than 1.00. The high self-satisfaction (HS) group was composed of the 17 subjects who received a score of 1.00 or more, while the low self-satisfaction (LS) group was composed of the 27 subjects who received a score of less than 1.00. The mean age for the HS group was 30.40 with a range of 19-38, while the mean age for the LS group was 27.30 with a range of 18-38. Both groups were matched on their immediate recall of a control passage, and the mean score for the HS group was 5.22 and for the LS group 5.12, a nonsignificant difference. Both groups were composed of chronic undifferentiated schizophrenics with only two chronic paranoid schizophrenics in the HS group and eight in the LS group.

Procedure

It is relevant to this experiment to note that two female assistants served in the various phases of the study. The recall phase of the experiment was conducted by one assistant, while the level of aspiration phase was conducted by the other assistant. Two different assistants were used in an effort to maintain some degree of independence between the different phases of the experiment proper.

Each subject was examined individually. After a brief general discussion, the subject was presented a control passage secured from the Wechsler Memory scale, Form 1 (1945, p. 6), to match the subjects on their immediate recall. Following the presentation of the passage, each subject rated the concepts AS, IS, and LLS on the semantic scale presented in counter-balanced order. Two matched groups of 17 HS subjects and 27 LS subjects, respectively, were secured for the experiment proper.

Recall series. This session was conducted 2-5 days after the initial phase of the study. For all subjects the following instruction was given:

Remember the ratings of yourself that you completed the last time? Well, as you know, they do reveal a lot of things about you. I have here an evaluation on what was revealed about you from the tests that you took. Of course, this will be strictly confidential. Listen very carefully while I read it to you because I want you to tell me as much about it as you can.

Following the reading of the passage, the subject was instructed to repeat as much of it as she could and was assured that it did not have to be in the exact words.

The experimenter read aloud the passages printed on a card. The experimenter practiced the reading prior to the experiment and found little difficulty in preserving uniformity from reading to reading.

The experimental passage. The experimental passage reproduced below has been subdivided into

items for scoring purposes. The passage was divided so that each item would include one idea, but that the recall of one item would not automatically include the recall of another.

You are an intelligent person/but you are unable to solve your problems./You are satisfied to do only enough to get by,/although you have the ability to do more./You do not always see things clearly,/even though you are capable of handling situations normally./You have good general knowledge,/and can assume responsibility./However, because you feel insecure,/you are afraid to try new things./You are able to get along with people,/but you are too easily offended./You could be self-sufficient,/but you prefer to be dependent upon others./

In order to note differences in the recall of types of items, each item was rated as "favorable" or "unfavorable." The ratings were made by the author and two other independent raters. The percentage agreement between the three independent raters was 84.30%. In the case of discrepancies, the final rating was decided upon after joint conference.

There were a total of 14 items, 7 favorable and 7 unfavorable.

Scoring. Each item recalled was assigned a weight of unity. An item was scored as a recall if reproduced accurately or if the idea itself was reproduced accurately. Inaccurate reproductions of items were not scored.

Level of aspiration series. Five days after the recall series, the Digit Symbol test secured from the Wechsler-Bellevue Intelligence Scale (Wechsler, 1944) was administered to each subject individually with the standard instruction provided by the manual, together with a 1 minute time limit. After the first performance, the following instruction was given: "You made a score of — on the test in 1 minute. Let's try it again. Here is another test which is done in the same way, but which uses different marks." An equivalent form of the Wechsler-Bellevue Digit Symbol test was used. After the subject completed the samples, the experimenter said: "How many of these do you think you will be able to do in 1 minute?" After recording the estimate, the subject was allowed to perform on the test.

Scoring. The deviation of estimate from performance was designated as the "D score." In each case the D score was the difference between the performance or actual score made and the estimate following it. When a subject estimated higher than the score she had earned, her D score was designated as positive. Whenever a subject estimated lower than the previous performance, her D score was negative. The D score provides not only a measure of the subject's aspiration but also of her adjustment to the reality of her own performance. A low D score implies somewhat better contact between goal and accomplishment.

A second D score utilized was the difference between the estimate just made and the performance following it. This measure reflects the level of success and failure in relation to the subject's goal setting.

Because of apparent skewness in the D scores, the results were analyzed by the nonparametric Mann-Whitney *U* test.

RESULTS AND DISCUSSION

Recall Series

In Table 1 the HS and LS groups are compared on the number of items recalled from the passage simulating a personality evaluation. Our data indicate that the HS group recalled significantly less items than the LS group, both in total recall and in the recall of items reflecting unfavorable personality characteristics. There was no significant difference in the recall of favorable items between the two groups. However, for the HS group alone, more favorable items were recalled than unfavorable items. It appears that, as predicted, the schizophrenic subjects with extremely high self-regarding attitudes tended to deny and suppress unfavorable personality features of themselves to such an extent that this was reflected in their performance. Since schizophrenic subjects with ex-

TABLE 1
COMPARISON OF MEAN RECALL SCORES

Type of item	HS recalls (N = 17)		LS recalls (N = 27)		<i>t</i>
	Mean	SD	Mean	SD	
Favorable	1.35	.84	1.78	1.13	1.31
Unfavorable	.94	.96	1.78	1.54	1.95*
Total	2.29	1.45	3.56	2.44	1.90*

* Significant at the .05 level, one-tailed test.

treme self-satisfaction tend to resist self-evaluation or any external promptings that may disrupt such self-definition they have adopted, the threat of the test situation undoubtedly contributed not only to the reduction in recall of unfavorable items but in the recall of favorable items as well. It is possible that, once developed such self-definition resist change and represents a prognostically poor sign for therapy. The question of therapy with such subjects invites further study.

Level of Aspiration Series

To begin with, the HS and LS groups were compared on their initial performance on the Digit Symbol test. The HS group obtained a mean score of 25.30 and the LS group a mean score of 26.00, a nonsignificant difference. In this respect, the two groups were matched in their performance on the Digit Symbol test, and this lessened the possibility of the effects of differential ability on the results of this phase of the study.

The discrepancy between performance and estimate (D score), with signs disregarded (i.e., positive or negative direction of estimates), provided a measure of each subject's adjustment to the reality of her own performance. The results of this analysis are given in Table 2. D scores secured from the discrepancies between Trial 1 and estimate revealed significantly higher D scores for the HS group as compared with the LS group. There were significantly greater discrepancies between the actual scores made on the first trial and

TABLE 2

MEAN DISCREPANCIES AND RESULTS OF APPLYING THE MANN-WHITNEY *U* TEST TO ARRAYS OF DISCREPANCY SCORES BETWEEN THE HS AND LS GROUPS

D score	HS group Mean	LS group Mean	<i>z</i> score
Trial 1 and estimate	9.71	3.60	2.61**
Estimate and Trial 2	9.71	4.96	2.37*
Trial 1 and Trial 2	3.06	3.67	.85

* Significant at the .05 level.

** Significant at the .01 level.

the estimates following it for the HS group as compared with the LS group.

D scores secured from the discrepancies between the estimates and subsequent performances (Trial 2) also revealed significantly higher D scores for the HS group as compared with the LS group. The discrepancies between the estimates made and the performances following it were significantly greater for the HS group than for the LS group. Analysis of the differences between Trial 1 and Trial 2 on the Digit Symbol test yielded no significant differences between the two groups.

Table 3 presents an indication of the percentages of subjects in the HS and LS groups in terms of the direction of their discrepancy scores. The D score between Trial 1 and estimate was designated as positive if the estimate following Trial 1 was higher, negative if lower, and zero if there was no change. The D score between estimate and Trial 2

TABLE 3

PERCENTAGE OF DIFFERENCES IN DIRECTION OF DISCREPANCY SCORES

	D score	HS group	LS group	Chi square	<i>p</i>
Trial 1 and estimate	Positive	35.29	33.33	5.03	.10
	Negative	58.83	33.33		
	Zero	5.88	33.33		
Estimate and Trial 2	Positive	64.71	62.96	3.19	.25
	Negative	35.29	22.23		
	Zero	0.00	14.81		
Trial 1 and Trial 2	Positive	47.06	74.07	4.52	.20
	Negative	23.53	18.52		
	Zero	29.41	7.41		

was designated as positive if Trial 2 following the estimate was higher, negative if lower, and zero if there was no change. Similarly, the D score was designated as positive if Trial 2 was higher than Trial 1, etc. Analyses of the percentages of directional differences by the chi square method revealed no significant differences between the two groups in the three conditions. That is, there were no significant differences between the HS and LS groups in the percentages of subjects showing positive or negative D scores.

The results of this phase of the study indicate that the differences between the HS and LS groups resulted not from the actual performance or in the direction of the D score, but in the setting of estimates or aspiration levels. Although there were no significant differences in the percentages of subjects revealing positive or negative D scores between the two groups, the HS group, as contrasted with the LS group, showed significantly larger discrepancy scores between their actual performance and estimate, whether in the positive or negative direction. The LS group, as contrasted with the HS group, was generally less extreme in setting their estimate either in the positive or negative direction. Since the discrepancy score represents the subject's adjustment to the reality of her own performance, the LS group was in better contact between goal and accomplishment than the HS group. In a sense, the LS group was capable of manifesting adaptive flexibility in such situations to a greater degree than the HS group.

Further comments appear indicated in regard to the composition of the HS group of this study, since it would seem anomalous for someone to evaluate his actual self higher than his ideal-self as did some of the schizophrenic subjects in the HS group. It would, indeed, be very unusual for someone to evaluate his actual self higher than his ideal-self, but such occurrences should be expected from some hospitalized schizophrenic patients. A frequently observed phenomenon in the clinical setting, are some schizophrenic patients revealing extremely high self-regarding attitudes who deny to themselves and to others

the extent of their pathology and discontent. Such schizophrenic subjects have an unusually high self-concept which, of course, represents an unrealistic self-appraisal. It should not be surprising, then, that some of the schizophrenic subjects comprising the HS group rated their actual selves higher than their ideal-selves. However, further study is needed to clarify the significance of such discrepancies reflected in the self-satisfaction index of such subjects.

SUMMARY

The present study sought to test two hypotheses: (a) schizophrenic subjects revealing extreme self-satisfaction tend to deny and suppress threatening features of themselves to such an extent that they will recall less items reflecting unfavorable personality characteristics from a passage designed to simulate a personality evaluation, than schizophrenic subjects admitting some dissatisfaction with themselves, and (b) schizophrenic subjects revealing extreme self-satisfaction will reveal greater discrepancy between their level of performance and level of aspiration than schizophrenic subjects not so characterized. Both hypotheses were supported when tested on a sample of 44 hospitalized schizophrenic women. Implications were drawn with regard to psychological adjustment.

REFERENCES

- ARIETI, S. *Interpretation of schizophrenia*. New York: Brunner, 1955.
- COWEN, E. L., HEILIGER, F., & AXELROD, H. S. Self-concept conflict indicators and learning. *J. abnorm. soc. Psychol.*, 1955, **51**, 242-245.
- MCQUITTY, L. L. A measure of personality integration in relation to the concept of self. *J. Pers.*, 1950, **18**, 461-482.
- OSGOOD, C. E., SUCI, G. J., & TANNENBAUM, P. H. *The measurement of meaning*. Urbana: Univ. Illinois Press, 1957.
- ROGERS, C. R. The significance of the self-regarding attitudes and perceptions. In M. S. Reymert (Ed.), *Feelings and emotions*. New York: McGraw-Hill, 1950.
- WECHSLER, D. *Measurement of adult intelligence*. Baltimore: Williams & Wilkins, 1944.
- WECHSLER, D. *A standardized memory scale for clinical use*. New York: Journal Press, 1945.

(Received September 12, 1960)

THE SCALING OF THE TAT FOR HOSTILITY BY A VARIETY OF SCALING METHODS

BERNARD I. MURSTEIN

Interfaith Counseling Center, Portland, Oregon

CHARLOTTE DAVID

Morningside Hospital, Portland, Oregon

DAVID FISHER

Tektronix, Incorporated

AND HANS G. FURTH¹

Catholic University

One of the key problems in the employment of projective techniques is the interpretation of the content of a TAT story with reference to possible behavioral correlates. For example, assume "A" and "B" both manifest an equal amount of hostility in their protocols to two different series of TAT pictures. Assume further that the hostility score employed is a sophisticated one taking cognizance of the inhibitions to aggression in the stories as well as to the direct expression of hostility. Does the equality of scores for A and B provide any foundation for the prediction that the behavioral correlates of their scores should be the same? Hardly, unless we can ascertain that the hostility cards in the series for each person are equal in their hostility-educing properties, and that the overall scores represent equal deviations from the stimulus properties of each card for each story. This latter statement is necessary since if A obtained his hostility score from telling chiefly hostile stories to nonhostile cards and nonhostile stories to hostile cards, the interpretative significance might differ from that attributed to B who followed the stimulus properties of the cards closely in giving hostile stories to hostile cards and nonhostile stories to nonhostile cards. Thus, the implications for interpretation may differ even though the overall scores are equivalent. This

topic has been dealt with more fully elsewhere (Murstein, 1961).

It is apparent that the relationship of responses on the TAT to the stimulus qualities of the cards may have important behavioral correlates which are helpful in the assessment of personality. Though several studies have been undertaken applying scaling techniques to thematic cards (Auld, Eron, & Laffal, 1955; Lesser, 1958) none have scaled the TAT in its entirety. Moreover, the previous studies employed only the Guttman technique. One might ask whether a series of scaling devices currently employed in measuring attitudes could be used to determine the stimulus properties of the entire series of TAT cards? If the stimulus value of the cards could be ascertained, then the relationship between the subject's response to the cards and the stimulus value might provide meaningful inroads into the study of personality. Our specific question therefore was formulated as follows: can a scale of hostility be constructed by each of the following scaling devices: Thurstone Equal Appearing Interval method (EAI); Successive Categories method (SC); Likert method; Edwards Scale Discrimination technique; and the Stouffer, Borghatta, Hays, and Henry H-technique?

PROCEDURE

Subjects were composed of 100 University of Portland undergraduate students, obtained from the various sections in general psychology. There were an

¹ The authors would like to thank Carol Bowdish, Walter Coulter, and Irvin Hansen for their contribution in the collection of data.

equal number of men (50) and women (50) for the Thurstone procedures. The Likert, Edwards, and Stouffer methods were applied to 42 men and 58 women. The same subjects participated in each scaling method with the exception that there were eight men missing and eight additional women present for the Likert, Guttman, Edwards, and H-technique procedures.

The EAI and SC methods differ primarily in that the former assumes that equal intervals exist for the 11 sorting categories while the latter method does not make this assumption but instead measures the actual width of the intervals between the cards.

The data for both methods, however, may be obtained in the same manner. Accordingly, groups of four to six subjects were instructed to stand before a table upon which was placed nine sheets of white paper, numbered one through nine, to represent nine categories of judgment. They were presented with the 31 TAT cards in random order and given these instructions:

You will be shown a series of 31 pictures which you are to judge objectively for the amount of hostility shown. By hostility I mean unfriendliness, anger, the desire to hurt either physically or mentally. The expression of hostility can vary from barely noticeable to extremely intense. It may be directed towards people, animals, objects, or nothing in particular. Your task is to judge the 31 cards according to the amount of hostility shown on each card. In front of you are numbers from one to nine which represent a continuum from the least amount of hostility to the greatest amount. Thus, the least hostile card would be put in pile Number 1 while the most hostile card would be put in pile Number 9.

Pile Number 5 represents the midpoint category separating the more than average hostile cards from the less than average hostile cards. For example, a card which seemed more hostile than average, but not extremely hostile might be put in a pile higher than the midpoint but yet not in one of the extreme piles.

Remember you are to judge these cards according to the amount of hostility they objectively possess, *not how you personally feel about them*. Do not forget to judge the blank card. Are there any questions?

The judgments were tabulated on a data sheet after each subject had completed the task. The data for the Likert, Edwards, and H-techniques were obtained in a group situation. With about 30 subjects seated in a classroom on each occasion, subjects were given the following instructions:

Now I would like you to look over each card individually and tell me how hostile it seems to be [hostility was defined in the same way as it was defined in the instructions for the aforementioned methods]. For each card I show you, check to see that you have the right number of the card as I call it out and then looking at the picture,

circle the phrase which best describes how hostile the picture appears. You have five choices: (1) very hostile, (2) fairly hostile, (3) undecided, (4) little hostility, (5) not hostile.

Slides containing facsimilies of the 31 TAT cards were presented to the three groups of approximately 30 students, one slide at a time. Each slide was projected on the screen for 45 seconds, during which time the subject recorded his judgment. To roughly control for serial position, one group received the pictures in numerical order, the second in inverse numerical order, and the last by starting with the middle numbered card and successively presenting the following cards in descending order from each side of the middle card. Thus, Card 11, the sixteenth card in the series of 31 cards, was presented first, followed by Card 10 (fifteenth card) and then by Card 12M (seventeenth card), etc.

RESULTS

Since there appeared to be little difference between the judgments of the sexes, the successive categories values for both sexes were correlated. The r of .93 indicated no reason why the judgments could not be viewed as coming from the same population of judgments. Accordingly, the median scale values (\bar{S}) for the EAI, SC, mean Likert values, and the EAI interquartile deviations (Q) for each card for the total group, are listed in Table 1 together with the Likert t values.

A test of internal consistency was applied to the values obtained via SC to determine whether the assumptions for scaling were supported. These assumptions were: (a) the projection of the cumulative proportion distributions for the various cards is normal on the psychological continuum; (b) the psychological dimension scaled is unidimensional; and (c) the standard deviations of the discriminial dispersions are equal. In using the χ^2 test, a fourth assumption made is that there is zero correlation among the stimuli, since the proportions used must be independent of each other (Edwards, 1957; Guilford, 1954). The χ^2 formula suggested by Guilford (1954, p. 232) was employed, which due to the large number of degrees of freedom, 210, was transformed into an approximate t ratio. The t value was 24.20 which was highly significant beyond the .001 level. It is apparent that one or more of the multiple assumptions made in scaling the cards was unjustified. To determine how important this finding was, we

TABLE 1

EQUAL APPEARING INTERVAL, SUCCESSIVE CATEGORY, Q , AND t VALUES FOR EACH TAT CARD

Card	Equal appearing interval scale value	Successive categories scale value	Equal appearing interval Q value	Likert mean value	t values for highest quartile vs. lowest quartile Likert judgment	p
12BG	1.13	.130	.31	1.11	2.00	.05
16	1.19	.190	1.07	1.20	.55	<i>ns</i>
10	1.27	.262	1.14	1.32	4.67	.01
8GF	1.41	.399	.69	1.91	1.94	.05
9BM	1.67	.580	1.00	1.18	2.45	.01
13G	2.31	.609	1.35	1.94	2.89	.05
14	2.80	.629	2.47	2.24	2.38	.05
17BM	2.89	.659	1.69	2.08	3.17	.01
2	3.00	.696	1.40	2.06	3.55	.01
13B	3.11	.734	1.39	2.13	5.45	.01
1	3.60	.896	1.61	3.28	2.23	.05
7GF	3.78	.947	1.66	2.65	2.54	.01
5	4.43	1.152	1.49	2.25	1.86	.05
7BM	4.56	1.192	1.47	3.97	2.00	.05
19	4.62	1.216	2.09	3.09	2.77	.01
6GF	4.76	1.263	1.52	3.95	4.53	.01
17 GF	4.77	1.266	2.52	3.50	2.75	.01
12F	5.23	1.424	1.85	3.78	3.74	.01
6BM	5.50	1.516	1.53	3.64	4.13	.01
20	5.56	1.538	1.81	3.59	2.36	.05
4	5.88	1.645	1.66	4.24	2.25	.05
8BM	5.90	1.650	2.49	3.37	3.78	.01
9GF	5.98	1.676	.69	3.86	2.71	.01
12M	6.11	1.763	1.89	3.96	2.11	.05
3BM	6.70	1.924	1.61	4.03	3.51	.01
11	6.79	1.968	2.38	3.29	1.53	<i>ns</i>
3GF	7.03	2.039	1.39	4.50	4.91	.01
18BM	7.82	2.436	1.06	4.75	2.90	.01
15	8.12	2.599	1.13	4.75	1.81	.05
13MF	8.15	2.616	1.17	4.77	3.75	.01
18GF	8.32	2.712	1.23	4.50	1.85	.05

obtained the size of the discrepancy between the theoretical and empirical proportions of judgments for each of the cards in the various categories. The theoretical proportions were obtained by taking the scale value of each of the 31 cards and subtracting it from each of the cumulative interval widths (Edwards, 1957). This yielded a 31×8 matrix of theoretical deviates with the columns representing the cumulative interval widths and the rows the various cards. By reference to the table of the normal curve these values were transformed into theoretical cumulative proportions. Each of these proportions based only on the knowledge of the interval widths

and the scale values of the cards was compared with its empirical counterpart. The average value for the discrepancy between the 248 theoretical and empirical proportions ($31 \text{ cards} \times 8 \text{ categories}$) was .038. This value is not exceedingly large, and indicates that the degree of lack of internal consistency in the scaling was not great although the confidence in the significance of the disparity is very high.

The Likert values were obtained by taking those individuals whose overall hostility score placed them in the top quartile and comparing their scores for each card with those persons whose score placed them in the lowest

TABLE 2
EQUAL APPEARING INTERVAL SCALE VALUES, Q VALUES,
AND t VALUES FOR NINE CARDS SELECTED FOR TEST
OF UNIDIMENSIONALITY

Card	Scale value	Q	t
10	1.27	1.14	4.67
13G	2.31	1.35	2.89
13B	3.11	1.39	5.45
7GF	3.78	1.66	2.54
6GF	4.76	1.52	4.53
9GF	5.98	.69	2.71
3GF	7.03	1.39	4.91
18BM	7.82	1.06	2.90
13MF	8.15	1.17	3.75

quartile. A *t* value was then obtained between the groups with regard to their scores on each card. Table 1 indicates that 11 of the 31 cards proved to be significantly differentiated at the .05 point, 18 at the .01 point, and only 2 proved to be not significant at all.

The Edwards Scale Discrimination Technique was utilized as follows:

1. The 15 cards above the median *Q* value were discarded.
2. The cards which were most representative of the entire range of \bar{S} values obtained from the EAI method were selected.
3. Those cards which, however, did not possess highly significant ($p < .01$) *t* values as obtained from the Likert method also were discarded.

TABLE 3
RESPONSES TO CARD 13MF AS RELATED TO
THE OVERALL SCORE FOR THE NINE CARDS

Total score	Card 13 MF		
	0	1	2
17-18	0	0	4
15-16	0	1	10
14	1	0	14
13	0	1	10
12	1	1	12
11	0	3	8
10	1	0	15
8-9	2	1	6
1-7	6	0	3
Sum	11	7	89

The result of this analysis was the selection of nine cards which were to be tested for unidimensionality via the coefficient of reproducibility. These cards from low to high stimulus value of hostility were 10, 13G, 13B, 7GF, 6GF, 9GF, 3GF, 18BM, and 13MF. The EAI, \bar{S} , *Q*, and *t* values for each of these cards are shown in Table 2.

The Stouffer, Borgatta, Hays, and Henry H-technique (1952) was used to determine the coefficient of reproducibility. The Likert responses (very hostile, fairly hostile, undecided, little hostility, not hostile) were collapsed into three judgments: hostile, undecided, and not hostile, which received weights of 2, 1, and 0, respectively. The distribution of total scores was then obtained for each of the nine cards, using the weights assigned to

TABLE 4
CUTTING POINT SELECTED FOR CARD 13MF
WHICH MEETS CRITERIA FOR SELECTION
VIA H-TECHNIQUE

Score for judgments of all nine cards	Nonhostile response 0, 1	Hostile response 2
≥ 10	9a	73
< 10	9	9a

Note.—a represents error cells. Card 13MF with cutting point of 0,1 and 2 fulfills the following conditions: (a) no error cell is greater than the smaller of the nonerror cell frequencies, and (b) the total error percentage is less than 30%.

the three response categories. Table 3 indicates by way of example, the relationship of the total response score on the nine cards to the score obtained for Card 13MF. Upon inspection of this table the best cutting points for the two possibilities 0, 1 vs. 2 and 0 vs. 1, 2 were selected, bearing in mind two criteria. There were (a) neither error cell has a higher frequency than the smaller of the two frequencies on the principal diagonal (nonerror), and (b) the sum of the frequencies in the two error cells should be less than 30% of the total frequency. An example of one of these splits for card 13MF is shown in Table 4.

By using more than one split with each card the nine original cards were condensed into four "contrived" cards. Each contrived card contained a "triplet," i.e., three cards

TABLE 5
TAT CARDS AND CUTTING POINTS USED IN THE CONSTRUCTION OF CONTRIVED CARDS
(*N* = 100)

Card	Response Weight		Frequency of hostile judgments	Contrived cards
	Negative	Positive		
18BM	0	1,2	93	— ^a
18BM	0,1	2	91	I
13MF	0	1,2	89	— ^b
13MF	0,1	2	82	I
3GF	0	1,2	81	I
9GF	0	1,2	77	— ^c
3GF	0,1	2	75	II
6GF	0	1,2	74	II
9GF	0,1	2	70	II
6GF	0,1	2	67	— ^c
7GF	0	1,2	55	— ^{c,d}
7GF	0,1	2	35	III
13B	0	1,2	35	III
13G	0	1,2	24	III ^c
13B	0,1	2	22	IV ^c
10	0	1,2	15	— ^{b,c}
10	0,1	2	9	IV ^c
13G	0,1	2	8	IV ^c

^a Card with this cutting point is closer to the end of the scale than is desirable.

^b Card not used because it appears with another cut in same contrived card.

^c Card with this cutting point has error cell with greater frequency than the smaller of the two frequencies on the principal diagonal.

^d Sum of both error cells greater than 30% of responses.

with prescribed cutting points which indicate which judgment or judgments are to be considered as a "hostile" choice, and which "non-hostile." Since there are two possible adjacent cutting points (0) vs. (1, 2) and (0, 1) vs. (2), each card could be used more than once if desired, using a different cutting point in each case. In order for a contrived card to be judged hostile, the responses to two or more of the members of the triplet must be judged hostile. With the number of possible scale types limited to five, the resulting coefficient of reproducibility was .965. This value is slightly inflated due to the fact that in choosing our cutting point we have taken advantage of favorable sampling errors. Nevertheless, the coefficient is sufficiently high to conclude that the responses to the cards can be reproduced with a satisfactorily high degree of accuracy from a knowledge of the total scores alone.

It should be noted that not all cards selected were able to fulfill Condition *a* mentioned above. Those cards not meeting this criterion, along with those not meeting other

criteria for selection, are listed in Table 5 along with the various cutting points, response category weights, the frequency (out

TABLE 6
FREQUENCY ASSIGNED TO EACH SCALE AND NONSCALE TYPE VIA H-TECHNIQUE

Response pattern	Frequency
+ + + + ^a	3
- + + +	0
+ - + +	1
+ + - +	0
+ + + - ^a	21
- - + +	0
- + + -	0
- + - +	0
+ - - +	1
+ + - - ^a	52
+ - + -	2
- - - +	0
- - + -	0
- + - -	3
+ - - - ^a	10
- - - - ^a	7
	100

^a Scale types.

of 100 subjects) with which the card in a particular split was judged hostile, and the contrived card into which the card was placed.

It is evident from examining this table that it was not possible to obtain a good split for a card which is judged hostile by 50% of the subjects. This failure is similar to that usually experienced with the Thurstone methods in attempting to get good differentiating items, or items with low Q values which at the same time represent the middle of the psychological continuum.

With four contrived cards there are 16 possible scores or types of response patterns, of which 5 may be designated as scale types and 11 nonscale types. The frequency with which each type was found is listed in Table 6. Examination of this table indicates that only 7 persons out of 100 are nonscale types.

Last, the values obtained for the 31 cards via the EAI, SC, and Likert methods were intercorrelated. The resulting correlations were: EAI vs. SC, .99; EAI vs. Likert, .94; and Likert vs. SC, .92.

DISCUSSION

Our data answer a few questions, and like much research, raise many others. Apparently, the TAT cards are readily scalable by a multitude of scaling methods. The Thurstone methods yielded a fairly representative range of the dimension of hostility with Q values not exceedingly higher than those often obtained for attitude statements. The coefficient of reproducibility of .965 obtained by the H-technique method also compares favorably with those usually obtained with attitude statements.

There are, however, further considerations. It is apparent from a perusal of Table 1 that the differential ability of the cards with regard to separating high hostility perceivers from low hostility perceivers is not greatly dependent upon the scaled value of the card. There are several instances where two cards are perceived as nearly equivalent on the dimension of hostility and yet one card is able to differentiate the aforementioned high and low hostility perceivers while the other is not. For example, Card 10 is given an EAI scaled value of 1.27 which is nearly identical to the scaled value of 1.19 received by Card 16. The

t value of Card 10, however, was 4.67 while that of Card 16 was .55. There are several possible explanations. One is that some cards contain many possible dimensions in their stimulus characteristics. High hostility perceivers (upper quartile in the Likert method), however, are more sensitive to the hostile possibilities of the card than to other motives such as achievement, sex, and affiliation, to name just a few. Low hostility perceivers (lower quartile in the Likert method), however, are probably either disposed to *avoid* seeing hostility or perhaps simply able to perceive the other dimensions as more strongly characterizing the picture. Card 10, which is described by Murray (1943) as "a young woman's head against a man's shoulder" (p. 19) would seem to be a multidimensional picture. There are many plausible explanations for the embrace, some positive, others negative. Card 16, however, is a picture devoid of any motives from a stimulus point of view. Hence, high hostility perceivers cannot choose any alternative except to perceive the picture as nonhostile. To do otherwise is to deviate sharply from the stimulus possibilities of the card. The low hostility perceivers likewise would be naturally expected to perceive little hostility.

It is thus conceivable that the number of alternative themes that can be perceived in a picture will determine its differential ability. The greater the number of possible themes in a card, the greater the differentiation between the judgments of persons high and low on one of the dimensions of the card.

It also follows that the greater the number of themes, the less likely a single motive is to receive a uniformly high judgment from all subjects. The reason for this assumption is that not all subjects will perceive the dominant motive since they may be more sensitized to other motives. The result is that the overall saliency of a motive is not only a function of the stimulus impact of the motive, but a function of the number of competing motives as well.

How then can one be sure as to which of these methods of judging is employed by the subject? One answer is to have individuals judge a picture for all possible motives. The perceptual ambiguity level could then be

determined (Kenny, 1961) by the equation $A = 1 - \sum p(i)^2$, where A equals the perceptual ambiguity of the picture and $p(i)$ equals the proportion of any i motive appearing in the picture.

A will be at a maximum when the proportions of all motives are equal (i.e., Motive A = .33, Motive B = .33, Motive C = .33; $A = .67$). A will decrease as the split between the proportion of two or more motives widens (i.e., Motive A = .50, Motive B = .45, Motive C = .05; $A = .45$). Current work on the differentiating value of the cards as a function of A is underway and will be reported in future articles.

Another problem with scaling procedures is that they usually have failed to incorporate the presence of inhibitory factors into the scaled values. If two people perceive an equal amount of hostility yet differ in the inhibitions expressed with regard to this hostility, their personalities might differ radically. Yet, our scaled values along with all other studies involving scaling of thematic stories would fail to take cognizance of this fact. It would seem, therefore, that the prediction of overt behavior from a knowledge of the scaled value of pictures might be improved if the scaled values reflected the multidimensionality of the stimulus properties of the picture. Perhaps some of the newer multidimensional scaling devices (Torgerson, 1958) may prove to be of greater value than the older methods.

The large variability of the Q values might tend to make one believe that Q might be regarded as an index of projection for a picture. Pictures with low Q values might be poor pictures for projective purposes while a high amount of variability for the *objective dimensionality* of a picture (high Q), might be considered a good index of projection. Little support, however, is given for the belief that high and low perceivers of hostility are differentiated by high Q variables if one examines Table 1. Two factors may serve to explain this result. First, a low Q value for the judgments of one dimension may become a high Q value when other dimensions are considered in the same picture. Secondly, a card may not be relevant for a given dimension. Accordingly, high Q values may result not be-

cause individuals differ as to the placement of the card on a dimension, but because they differ as to whether or not the picture belongs in the dimension continuum at all. Under these circumstances, a high Q may merely reflect a high amount of error variance in the judgment due to the random assignment of values to a card when it does not seem applicable to a dimension. Thus, for Q to be regarded as an accurate measure of projection we should ascertain that the card is relevant to the dimension being judged, and that it taps this dimension *only*. Since it is doubtful that the current TAT cards meet both criteria completely, the use of Q as a projective index would seem to have some drawbacks.

There are obviously a good many difficulties in the scaling of thematic cards and the question may arise as to whether it is worth all of the trouble. Is a knowledge of the stimulus value important? Cannot a skilled clinician "get along" without knowing the stimulus value of the cards? It is our belief that behavior may be viewed as the pooled interaction of stimulus, background, and organismic variables, a view very close to that expressed by Helson (1955). From this frame of reference, knowledge of the stimulus properties seems essential to the accurate prediction of behavior. The "good" clinician probably carries in his head a normative index of responses to each of the TAT pictures which serves as a rough estimate of the stimulus value. But, the clinician limited by his own experience probably could not achieve the accuracy of estimation that the actual measurement of a well standardized group would achieve.

Yet another important factor is the determination of the relationship between the stimulus properties of a card, the response elicited, and the possible behavioral correlate. This important area has been untouched by psychologists because of a lack of knowledge of the stimulus properties of the cards (Murstain, 1959). It should be possible shortly, to determine the relationship, if any, between perceptual deviancy and behavioral deviancy. It is not assumed that maladjustment is a simple linear function of the discrepancy between the stimulus properties of a picture and the story told to it. These may be a curvi-

linear, hyperbolic, parabolic relationship, or perhaps, no relationship. The determination of an answer to this important question only awaits our quantification of the stimulus properties of our projective instruments.

In closing, it should be emphasized that the values obtained may not hold for other kinds of students at other locales. In fact serial position effects with each scaling method as well as between the methods have not been adequately controlled. To do so would have involved a great number of subjects, which, while desirable, was not practical, not directly pertinent to the rather broad purpose of this study. Such refinements should, however, be utilized where the scale values themselves are of concern rather than the question of whether scaling itself can be achieved.

SUMMARY

The purpose of this study was to determine whether the entire set of 31 TAT cards could be scaled for the dimension of hostility through the use of several widely used scaling methods.

A group of 100 undergraduate psychology students were administered the TAT cards via slide projections and asked to judge the slides with regard to the dimension of hostility. The judgments were scaled by the Thurstone Equal Appearing Interval and Successive Category methods, the Likert method, Edward Scale Discrimination technique, and Stouffer, Borgatta, Hays, and Henry H-technique. By employing various criteria such as adequate range coverage, and differential ability between high and low hostility perceivers, eight cards were finally selected. These were 10, 17BM, 7GF, 6GF, 9GF, 3GF, 18BM,

13MF. The coefficient of reproducibility for these cards using the H-technique method of "contrived cards" was .96. It was concluded that all of the aforementioned methods could be used in scaling the dimension of hostility. The implications of the results with regard to future work in the area of personality were discussed.

REFERENCES

- AULD, F., ERON, L. D., & LAFFAL, J. Application of Guttman's scaling method to the TAT. *Educ. psychol. Measmt.*, 1955, 15, 422-435.
- EDWARDS, A. L. *Techniques of attitude scale construction*. New York: Appleton-Century-Crofts, 1957.
- GUILFORD, J. P. *Psychometric methods*. New York: McGraw-Hill, 1954.
- HELSON, H. An experimental approach to personality. *Psychiat. res. Rep.*, 1955, 2, 89-99.
- KENNY, D. T. A theoretical and research appraisal of stimulus factors in the TAT. In J. Kagan & G. Lesser (Eds.), *Contemporary issues in thematic apperceptive methods*. Springfield: Charles C Thomas, 1961. Pp. 288-310.
- LESSER, G. S. Application of Guttman's scaling method to aggressive fantasy in children. *Educ. psychol. Measmt.*, 1958, 18, 543-551.
- MURRAY, H. A. *Thematic Apperception Test manual*. Cambridge: Harvard Univer. Press, 1943.
- MURSTEIN, B. I. A conceptual model of projective techniques applied to stimulus variations with thematic techniques. *J. consult. Psychol.*, 1959, 23, 3-14.
- MURSTEIN, B. I. The role of the stimulus in thematic apperceptive methods. In J. Kagan & G. Lesser (Eds.), *Contemporary issues in thematic apperceptive methods*. Springfield: Charles C Thomas, 1961. Pp. 229-273.
- STOUFFER, S. A., BORGATTA, E. F., HAYS, D. G., & HENRY, A. F. A technique for improving cumulative scales. *Publ. opin. Quart.*, 1952, 16, 273-291.
- TORGERSON, W. S. *Theory and methods of scaling*. New York: Wiley, 1958.

(Received September 14, 1961)

THERAPISTS' JUDGMENTS CONCERNING PATIENTS CONSIDERED FOR PSYCHOTHERAPY¹

SOL L. GARFIELD AND D. C. AFFLECK

University of Nebraska College of Medicine

In recent years, an increasing amount of attention has been devoted to the problem of duration of stay in psychotherapy. Reports from many diverse types of clinical settings have indicated that early discontinuation in outpatient psychotherapy is a reliable finding of some importance (Affleck & Mednick, 1959; Garfield & Kurz, 1952; Rogers, 1960). Several studies have attempted to appraise selected patient variables related to and predictive of continuation in psychotherapy (Rosenthal & Frank, 1958; Rubinstein & Lorr, 1956; Sullivan, Miller, & Smelser, 1958; Taulbee, 1958). With the possible exception of educational level, the findings on most of these variables have been inconsistent. Research in our setting on patient attributes related to termination indicated that diagnosis, sex, age, and education were not significantly related to duration of stay (Garfield & Affleck, 1959). Education below a certain minimal point may be important, but we found no evidence to indicate its usefulness as a predictor with persons who have gone beyond the eighth grade in school.

The general failure to relate broad patient variables to attrition led to an interest in the therapist as a variable affecting attrition rates. While it is apparent that the interaction between the individual patient and the individual therapist is exceedingly important for the problem of attrition, we were interested first in getting a better understanding of the orientation that therapists have toward candidates for therapy in general. Are there common points of view toward patients? What patients are initially viewed in a highly favorable way? For what reasons is this the case? Which patients are seen negatively?

Are anxiety and defensiveness related to the judgments and attitudes therapists have toward patients? These were some of the questions that led to an initial exploratory study of therapists' attitudes toward therapy candidates. This in turn was part of a larger study of variables related to continuation and progress in psychotherapy.

PRESENT STUDY

In this investigation, therapists were asked to complete a brief questionnaire and checklist at staff meetings at which cases were discussed and considered for outpatient psychotherapy. The questionnaire included open-ended questions on assets, deficiencies, goals in therapy, and likely problems in therapy. Each therapist also was asked to rate each patient on a four-point scale in terms of therapeutic prognosis—excellent, good, fair, or poor. Similar ratings were requested concerning the degree of anxiety in the patient, the latter's defensiveness or rigidity, the rater's personal feelings toward the patient, and the rater's interest in taking the patient on for psychotherapy.

The ratings were secured in a regular outpatient staff meeting which met weekly for 2 hours. Two to three cases were discussed at each meeting. These cases had been seen previously by a psychiatric resident and social worker, and in about one-half of the cases by a psychologist. The intake reports were all read in their entirety. After the intake material was presented to the staff, but prior to any discussion of the case, each of the individuals at the staff meeting was asked to fill out the questionnaire.

Responses were secured from 20 different therapists from three disciplines: psychiatry, clinical psychology, and psychiatric social work. The number of patients rated and

¹ Presented in part at the Annual Meeting of the American Psychological Association, Chicago, September 1960.

TABLE 1
CORRELATIONS BETWEEN RATED PATIENT VARIABLES

Therapist	N	Correlation coefficient							
		A-B	A-C	A-D	A-E	B-C	B-D	C-D	D-E
1	23	.03	-.59(19)**	.72**	.41*	.15(19)	.09(22)	-.33(18)	.61(22)**
2	25	.18	-.62(24)**	.49(24)*	.21(24)	-.07	.35	-.42*	.62**
3	26	.44*	-.70**	.55**	.51**	-.07	.53**	-.39*	.96**
4	28	.33	-.54**	.66**	.76**	-.14	.15	-.26	.70**
5	18	.66(16)**	-.50(13)	.69(17)**	.89**	.02(13)	.50*	-.50(14)	.60**
6	17	.26	-.91(16)**	.82**	.77**	-.18(16)	.46	-.65(16)**	.78**
7	27	.38*	-.26	.76**	.65**	.16	.29	-.28	.66**
8	12	.42	-.10	.45	.32	-.13	.41	-.76**	.63*
9	20	.51*	-.33	.74**	.73**	.00	.52	-.45*	.95**
10	31	.68**	.13	.60**	.70**	.25	.64**	-.02	.58**
11	17	.40(16)	-.04(15)	.51*	.55*	-.15(14)	.25(16)	-.54(15)*	.43
12	29	.05	-.53**	.73**	.26	-.20	.22	-.18	.67**
13	14	.28	-.56*	.54*	.71**	.10	-.14	-.76**	.54*
		.38	-.53	.66	Median Correlation				
					.65	-.07	.35	-.42	.63

Note.—Numbers in parentheses indicate number of cases when they vary from that indicated in Column 2.

A—Therapeutic prognosis.

B—Degree of anxiety.

C—Defensiveness and rigidity.

D—Personal feelings.

E—Interest in taking into treatment.

* Significant at .05 level.

** Significant at .01 level.

evaluated by each therapist varied from 7 to 32 with a median number of 18 patients. All of the patients were individuals who had applied for outpatient psychotherapy and had been recommended for intake evaluation by the initial screening committee. The patients consisted of 18 men and 20 women ranging in age from 13 to 51 years, with a median age of 27 years. In terms of diagnosis, the group was as follows: Psychoneurosis, 16; Personality Disorders, 14; Psychosis, 3; and other diagnoses, 5.

RESULTS

The responses secured from the therapists were tabulated for each category of response. In order to evaluate the reliability of the ratings, average intercorrelations were computed on the five raters who had seen at least 16 patients in common. Two of these raters were staff psychologists, one was a staff psychiatrist, and two were psychiatric residents. Ebel's (1951) technique for estimating the reliability of ratings was used. The judges showed a high degree of agreement in their ratings of therapeutic prognosis ($r = .88$),

personal feelings toward the patient ($r = .79$), interest in taking the patient on for therapy ($r = .80$), and patient's anxiety level ($r = .88$). Moderate agreement was obtained in the judges' estimates of the patient's defensiveness ($r = .68$).

The ratings obtained were then intercorrelated where appropriate. Thirteen therapists who rated at least 12 patients were used in this analysis, which forms the first part of our report. We shall discuss now the ratings of therapeutic prognosis and their relationship to other judgments.

In line with other findings, it was hypothesized that degree of anxiety would be correlated positively with prognosis whereas defensiveness and rigidity would be negatively correlated with prognosis (Rubinstein & Lorr, 1956; Taulbee, 1958). These predictions were generally supported, although not to a marked degree. As can be seen in Table 1, all the correlations between therapeutic prognosis and degree of anxiety are positive, but less than a third of them are statistically significant, with the median being .38. Ratings of anxiety thus bear only a limited relationship

to ratings of prognosis. The relationship between prognosis and defensiveness appeared to be somewhat more marked with approximately half of the correlations approaching significance. As might be anticipated, this relationship was negative in all but one case. Apparently, our judges react somewhat more strongly to defensiveness and rigidity in relation to prognosis than they do to anxiety in this regard.

Ratings of prognosis, on the other hand, were highly correlated with positive feelings of the judges toward the patient. Only one of the correlations was not significant at least at the .05 level of confidence, with the median correlation being .66. The personal feeling of the raters thus appears most closely related to ratings of prognosis, or vice versa. Ratings of "interest in taking the patient into treatment" were also highly correlated with both ratings of therapeutic prognosis and the personal feelings of the raters toward the patients. The latter finding suggests that personal feelings toward the patient, interest in taking the patient on for therapy, and judgments of prognosis may be manifestations of the same positive view of the patient. One cannot state whether the raters "like" patients with good prognosis, or whether a good prognostic rating is given to patients that the therapist reacts to personally in a positive fashion.

The other findings were not as marked, although there was some negative relationship between the personal feelings of the rater and defensiveness of the patient. It would thus appear that judgments of prognosis are most closely related to the personal feelings of the therapist judges (or vice versa), and that the latter bear more relationship to judgments of defensiveness and rigidity than they do to judgments about the patient's anxiety. This pattern generally is congruent with that recently reported by Strupp and Williams (1960). In studying two therapists, they found that "nondefensive, insightful, likable and well-motivated patients were seen as most likely to improve in psychotherapy" (p. 440).

ASSETS FOR PSYCHOTHERAPY

As mentioned previously, each rater also was asked to list the therapeutic assets for

TABLE 2
PATIENT ASSETS LISTED FOR PSYCHOTHERAPY

Asset	Frequency
Intelligence	138
Anxiety-discomfort	112
Motivation	98
Age	49
Insight-awareness of problem	49
Past adjustment	29
Ability to relate	20

each patient as well as to indicate likely problems to be encountered in psychotherapy. A total of 532 responses pertaining to patient assets were obtained with a variable number being listed for any given case. The average number per patient was one and a half. After a preliminary analysis was made of all the individual responses, the results were grouped into appropriate categories. Although a very large number of responses were listed, these could be classified with little difficulty into a relatively small number of categories. All of the items which were mentioned at least 10 times are presented in Table 2.

As noted in Table 2, three categories make up over half of the listed assets, i.e., intelligence, anxiety, and motivation. When age and insight are added, these five account for over 80% of all the assets listed for these patients. On the basis of such ratings, one might infer that the average therapist prefers a patient who is intelligent, anxious, well motivated for therapy, young, and with some insight into his difficulties! This seems to be borne out by an analysis of the assets listed for patients in relation to the ratings by our judges of personal feelings toward the patients. When the total group of patients is dichotomized in terms of the median ratings on this variable, it is noted that the group which receives the higher ratings also receives almost twice the frequency of listed assets. This difference is significant at the .01 level of confidence ($\chi^2 = 15.42$, $df = 1$). With the exception of age, the assets mentioned are linked more frequently with patients given high personal preference ratings by the therapists. The preferred therapy patient, as inferred from these listings by our sample of therapists, bears a close resem-

TABLE 3

DIFFERENCES BETWEEN MEAN SCALE RATINGS OF
12 TERMINATORS (T) AND 12 REMAINERS (R)

Scale	Mean of T	Mean of R	t
Defensiveness	1.88	1.84	.28
Anxiety	1.85	1.85	—
Prognosis	3.00	2.64	1.88*
Personal feelings	2.64	2.49	.97
Interest in taking	2.69	2.52	.87

* Significant at .05 level of confidence, one-tailed test.

blance to the type of preferred patient mentioned in the research by Hollingshead and Redlich (1958).

It is of interest also to comment on the variability with which the various assets were listed by different therapists. Intelligence, for example, was listed in one out of eight cases by one person, but in almost two out of three cases by another. In a similar fashion, anxiety or discomfort was given as an asset in over half of the cases by one therapist, but in only 1 case out of 16 by another. Thus, while there is some consensus concerning desirable features in a psychotherapy patient, there is some variation among therapists concerning the frequency of emphasis on certain aspects of the patient. Our data are too meager at this point to permit us to infer any particular relationship between the pattern of assets listed by specific therapists and other judgmental variables.

THERAPISTS' JUDGMENTS AND DURATION OF STAY IN PSYCHOTHERAPY

Of the 38 patients who were evaluated initially at the outpatient staff conferences, 24 were assigned to a therapist in our outpatient clinic, thus allowing for some follow-up study. All of the therapists who saw these patients participated in the initial rating procedures. The other 14 patients were referred to other agencies, clinics, or hospitals. In a few of these cases, no treatment was recommended. The median number of interviews kept for the group of patients assigned to therapy here was 17. (This atypically high figure may be somewhat misleading. There were 11 patients

who kept 12 or less interviews, a value which is the same as that previously reported as the median on a much larger sample of patients—Garfield & Afleck, 1959.)

It was hypothesized that ratings of low defensiveness, high anxiety, good prognosis, positive personal feelings, and a positive interest in taking would be related to greater duration of stay in psychotherapy. Differences between the mean ratings of all scales for patients above and below the median were analyzed. Each patient was rated by a median of 9 raters, with a range of 6 to 14 raters. Table 3 presents the results of these analyses.

The only rating which was significantly related to duration of stay was prognosis; patients remaining in therapy longer are rated initially as having a better prognosis. Despite the moderate intercorrelation of prognosis with the personal feelings of the therapist, ratings of the latter were not significantly related to duration of stay. The failure of ratings of personal feelings, interest in taking the patient on for therapy, anxiety, and defensiveness to predict duration of stay is of interest in the light of our previous findings on therapists' preferences. Tentatively, it appears that the set therapists develop toward patients on these dimensions are reliable, but have no predictive validity as regards duration of stay. While interest in taking a patient on for therapy and the personal feelings of the therapist toward the patient were significantly correlated with ratings of prognosis and thus suggestive of a common view of the patient, only the latter appeared related to duration of stay in psychotherapy.

SUMMARY AND CONCLUSION

Our findings show a high degree of agreement among therapists in terms of judgments of prognosis, personal feelings toward patients, and interest in taking patients on for psychotherapy. Ratings of these variables also show moderate intercorrelations. This suggests that certain patients have a high valence for therapists and that there is agreement among therapists as to who these patients are. A somewhat similar consensus was evident in the listing of patient assets for psychotherapy where five assets constituted 80% of the total listed. In terms of the assets listed and the

negative feeling expressed by therapists toward defensiveness in patients, it would appear that a positive reaction is expressed toward the patient least difficult to work with and, possibly, the person least in need of skilled help. It was further demonstrated that patients who evoke positive feelings from therapists are characterized by those therapists as having significantly more assets, particularly intelligence, motivation, anxiety, and insight.

When the ratings were related to actual duration of stay, it was found that patients remaining in therapy longer were rated as having a better prognosis. None of the other ratings were significantly related to duration of stay. While therapists show high agreement in their preferences and personal feelings for patients, these ratings were not related to actual duration of stay.

REFERENCES

- AFFLECK, D. C., & MEDNICK, S. A. The use of the Rorschach test in the prediction of the abrupt terminator in individual psychotherapy. *J. consult. Psychol.*, 1959, 23, 125-128.
- EBEL, R. L. Estimation of the reliability of ratings. *Psychometrika*, 1951, 16, 407-424.
- GARFIELD, S. L., & AFFLECK, D. C. An appraisal of duration of stay in outpatient psychotherapy. *J. nerv. ment. Dis.*, 1959, 129, 492-498.
- GARFIELD, S. L., & KURZ, M. Evaluation of treatment and related procedures in 1,216 cases referred to a mental hygiene clinic. *Psychiat. Quart.*, 1952, 26, 414-424.
- HOLLINGSHEAD, A. B., & REDLICH, F. C. *Social class and mental illness: A community study*. New York: Wiley, 1958.
- ROGERS, L. S. Drop-out rates and results of psychotherapy in government aided mental hygiene clinics. *J. clin. Psychol.*, 1960, 16, 89-92.
- ROSENTHAL, D., & FRANK, J. D. The fate of psychiatric clinic outpatients assigned to psychotherapy. *J. nerv. ment. Dis.*, 1958, 127, 330-343.
- RUBINSTEIN, E. A., & LORR, M. A. A comparison of terminators and remainers in outpatient psychotherapy. *J. clin. Psychol.*, 1956, 12, 345-349.
- STRUPP, H. H., & WILLIAMS, JOAN C. Some determinants of clinical evaluations of different psychiatrists. *AMA Arch. gen. Psychiat.*, 1960, 2, 434-440.
- SULLIVAN, P. L., MILLER, C., & SMELSER, W. Factors in length of stay and progress in psychotherapy. *J. consult. Psychol.*, 1958, 22, 1-9.
- TAULBEE, E. S. Relationship between certain personality variables and continuation in psychotherapy. *J. consult. Psychol.*, 1958, 22, 83-89.

(Received September 19, 1960)

AN EMPIRICAL SCALE OF THERAPIST VERBAL ACTIVITY LEVEL IN THE INITIAL INTERVIEW¹

EDMUND S. HOWE AND BENJAMIN POPE

University of Maryland School of Medicine

Subjecting the psychotherapist to examination as an independent variable reflects acceptance of the proposition that, regardless of his theoretical orientation, what the therapist says is of central importance in the therapeutic transaction. Subjecting him to similar examination in the initial interview implies that the therapist's mode of verbalization may have an important bearing upon achievement of his diagnostic or other goals.

The last 20 or more years have seen two major transitions in the tactics and strategy of the initial interview, which have arisen largely as influences of Freudian psychoanalytic theory. On the one hand there has been increased understanding that, no less than the formal therapeutic interview, the initial interview involves an interpersonal process influencing both the patient and the therapist as a participant observer. On the other hand, simultaneously, there has been increasing departure from the "fact gathering" typical of the earlier psychiatric interview, to a process in which the patient is encouraged, through relative passivity on the part of the therapist, spontaneously to unfold his story as he himself feels it. Thus, many contemporary writers

(e.g., Deutsch & Murphy, 1955; Finesinger, 1948; Gill, Newman, & Redlich, 1954) attempt to arrive at some kind of working diagnostic formulation during an initial interview not by eliciting a mass of factual information about various sectors and stages of the patient's life history; but instead by following the patient's own leads, his sequential account of himself, his life, and his difficulties.

These transitions in the form of the initial interview can be described in terms of increasing adoption of the *projective* interview, in which it is now commonly accepted that one is apt to discover more information of a relevant nature either by remaining silent, or at most by asking rather vague, nonleading questions onto which the patient may project his own referents, and his own interpretation of what is "meant." In this way one learns much not only about circumstantial (factual) material, but also about those contiguous motivational and associational processes which usually lie nearer to the heart of the matter.

The foregoing developments have given rise to the concept of Therapist Activity Level (e.g., Finesinger, 1948) with the attendant implication that lower Activity Levels are potentially more advantageous than higher ones, for the purposes of gathering relevant information, fostering the development of transference reactions, and avoiding a shift into a social or personal relationship with the patient. (There are, however, obvious exceptions to any general rule, such as the use of higher levels of activity for such supportive purposes as encouraging the inhibited patient to talk during an initial interview [Gill et al., 1954] or to prevent acting-out behavior.)

These commonly assumed benefits of maintaining a low level of verbal activity remain hypothetical, however, since they have never been subjected to experimental scrutiny. This

¹ This paper arises out of research supported by Pilot Evaluation Grant No. 2M-6408 from the National Institute of Mental Health of the National Institutes of Health, United States Public Health Service. The late Jacob E. Finesinger was the principal investigator. Thanks are acknowledged for his continuous encouragement and wholehearted support of this work until his untimely death in June 1959. A paper based partly upon the first four studies was presented by Pope, Howe, and Finesinger to Division 12 at the Annual Convention of the American Psychological Association in Cincinnati, Ohio, September 1959. Completion of Study 5 and of the present manuscript have been facilitated during tenure by Edmund S. Howe, of Research Grant M-3355, also from the National Institute of Mental Health.

paper constitutes a preliminary basic step in a research program the aim of which is to evaluate the role played by, and the impact upon the patient of, the therapist's Activity Level in the initial interview. The experiments to be reported at this time were performed (a) to examine the rateability of the concept of Activity Level in terms of three assumed attributes (to be discussed below); (b) to develop an Activity Scale for subsequent measurement procedures; and (c) to explore some of the empirically controllable variables that might affect the reliability of application of such a scale to actual interview material.

The choice of a definition of Activity, however, presents a problem, for its attributes are not clear, and have never been spelled out. Deutsch and Murphy (1955) for example, made no attempt to define Activity, other than implicitly by rejecting the question-and-answer interview pattern, and instead proposing a "process of facilitation through the selective repetition in interrogative form of the patient's remarks" (p. 18). Finesinger (1948) likewise skirted the conceptual problem of definition in expressing his preference for Activity which is kept "as low as is consistent with the attainment of therapeutic plans and goals" (p. 192).

Several research workers (e.g., Bordin, 1955; Dibner, 1953; Osburn, 1951) have accepted the term *Ambiguity* as a significant aspect of therapist behavior. Dibner in fact showed that certain consequences in the patient's behavior (e.g., increased "anxiety") follow greater therapist *Ambiguity*. Bernstein, Lennard, and Palmore (1958) likewise observed greater "ease of communication" by the patient following greater therapist *Specificity* (i.e., less *Ambiguity*). Several years earlier Snyder (e.g., 1945) investigated *Lead*, which he assumed to be a primary dimension of therapist verbal behavior. (Indeed, it is interesting to note that when Freud [1948] himself abandoned hypnosis in favor of the psychoanalytic technique, he contrasted the suggestive nature of the former with the non-leading character of the latter.) Finally, it is considered that a therapist response may also be looked at from the standpoint of the degree of *Inference* which it carries, or which it conveys to the patient. In the studies to be

described these three attributes, *Ambiguity*, *Lead*, and *Inference*, will be used to characterize what is meant by variations in Activity Level. It was assumed for the purpose of these studies that the three attributes are moderately (if not highly) intercorrelated, so that the three terms are to some extent interchangeable. Thus, *Ambiguity* subjectively feels as though it would be negatively correlated with *Lead* and with *Inference*, whereas the last two would be positively correlated with each other. To this extent Activity is assumed, for present purposes, to be one-dimensional.

METHOD

Study 1. A broad variety of over 20 published psychotherapy interviews involving different types of patients, different phases of treatment, and therapists of different theoretical allegiance, were used as source material to compile a representative sample of 50 abstract descriptions of therapist verbal responses. Thirty Board-certified psychiatrists rated each of these descriptive responses (presented on individual 3 × 5-inch cards) for Activity Level along an 11-point scale. A broad working definition characterized Activity Level as follows:

A high-active response from the therapist is not, of course, necessarily one which has greater length. It does, however, have relatively *low Ambiguity* about it; it involves a *marked degree of Lead* by the therapist; and it carries a *high degree of Inference*. Conversely, a low-active response is *highly ambiguous*; it manifests a *low degree of Lead* by the therapist; and it carries a *low degree of Inference*. Thus, compare the following three descriptive responses:

1. Therapist gives a general, unfocussed invitation for the patient to talk.

2. Therapist asks the patient to describe the last occasion when a pattern of symptoms occurred.

3. Therapist explores the patient's feelings about something just reported by the patient.

Going from 1 to 2 through 3, the responses become less ambiguous, they show progressively more lead, and they connote an increasing degree of inference. . . .

Each rater also sorted a duplicate set of cards into one of three groups: (a) responses primarily or mainly diagnostic in purpose (ignoring secondary, therapeutic value); (b) responses primarily or mainly therapeutic in purpose (ignoring secondary, diagnostic value); and (c) responses fitting neither category. The two tasks were given in one of two sequences to alternate subjects.

Results of Study 1. A Lindquist (1953, pp. 267-273) Type I analysis of the 11-point rating data established (a) an overall difference among the 50

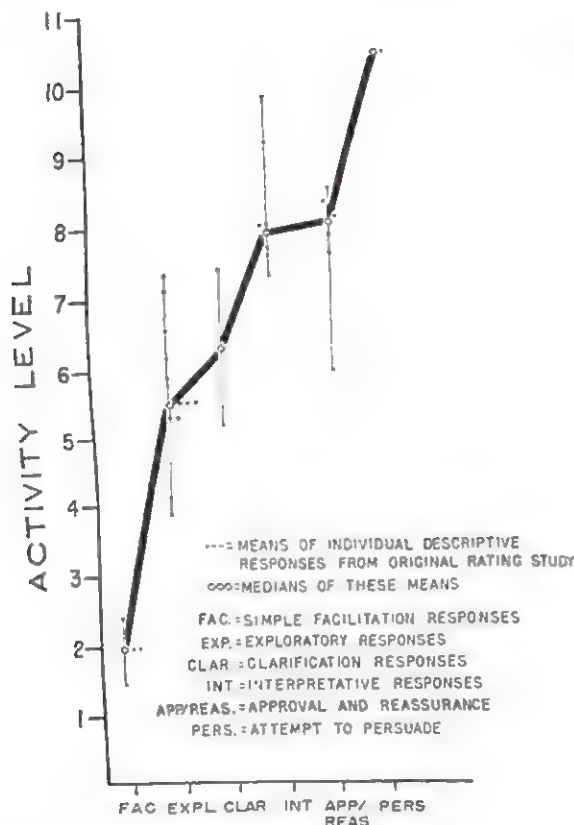


FIG. 1. Mean and median Activity Levels of 35 therapist responses placed (with significant statistical agreement among 10 subjects) into five conventional categories of therapist behavior.

item means ($p < .001$); (b) a nonsignificant Sequence main effect; and (c) a nonsignificant Sequence \times Items interaction effect ($p > .05$). Since the Activity ratings were thus clearly not altered as a result of prior judgments of diagnostic vs. therapeutic value, the rating data were then pooled. The interclass r was .50; the reliability of average ratings, .93 (Guilford, 1954).

After computation of appropriate chance frequencies via Fisher's Exact test (Siegel, 1956) it was established that the 30 subjects significantly agreed upon only 10 of the responses as being primarily of "treatment value," and upon only 7 as being primarily of "diagnostic value." A comparison of the mean Activity Levels of these two groups of responses via the Mann-Whitney U test (1947) showed the responses in the treatment category to be more active ($p < .001$). This accords with commonsense expectation, and constitutes a modicum of face validity for the working concept of Activity.

Study 2. This was undertaken to examine the relationship between Activity Level as rated in Study 1 and five conventional labels frequently applied, in the contemporary literature on psychotherapy research, to various categories of therapist operations. Ten new subjects, four psychiatrists and six clinical

TABLE 1
MEAN ACTIVITY LEVEL ASSIGNED 35 THERAPIST RESPONSES BY PSYCHIATRISTS, AND BY MEDICAL STUDENTS WITH HIGH AND LOW INTEREST IN PSYCHIATRY

Type of Subject	N	Mean	σ
Psychiatrists	15	6.17	3.03
Students with high interest	19	6.44	3.02
Students with low interest	18	7.08	3.17

psychologists with significant therapeutic experience, sorted the set of 50 therapist responses into the following six categories: Simple Facilitation; Exploration; Clarification; Interpretation; Approval and Reassurance; and Unclassifiable. Subjects were given a broad working definition of each category. For example, Simple Facilitation responses were defined as "quite unfocused responses designed to get the patient talking, and to keep him talking, without imparting any direction to him whatsoever."

Results of Study 2. Application of Fisher's Exact test (Siegel, 1956, Table D) to each category showed better than chance agreement among subjects on 39 of the 50 responses, only 4 of which were in the Unclassifiable category. A Kruskal-Wallis (1952) one-way analysis of variance of the Activity Levels of the 35 responses within the five conventional categories yielded an H of 22.79 (with $df = 4$, $p < .001$). Figure 1 shows the data plotted, Activity Level medians against type of conventional category. The relative order of magnitude of median Activity Level for the categories largely accords with subjective commonsense expectation. The "Persuasive" response, which as one might expect is rated most Active, was not actually included in Study 2, since it was the only one of its kind in the original set of 50 responses. It is presented in Figure 1, however, for the sake of perspective and completeness over the entire range of therapist operations actually studied. It is noteworthy that there is considerable overlap between adjacent types of conventional categories of

TABLE 2
RANK ORDER CORRELATIONS BETWEEN THE ACTIVITY RATINGS OF 35 THERAPIST RESPONSES BY PSYCHIATRISTS AND BY MEDICAL STUDENTS WITH HIGH AND LOW INTEREST IN PSYCHIATRY

Types of Subjects	rho
High interest vs. psychiatrists	.867***
High interest vs. low interest	.737***
Low interest vs. psychiatrists	.636***

*** $p < .001$.

TABLE 3
PARALLEL ACTIVITY SCALES A AND B, AND MEAN ACTIVITY LEVEL OF EACH PAIRED ITEM

Scale A		Scale B	
Mean AL	Descriptive responses	Mean AL	Descriptive responses
1.7	Therapist uses a single word or syllable to give the patient an invitation to continue.	1.6	Therapist says "Hm-hm" to convey acceptance and understanding of the patient.
3.3	Therapist repeats exactly what the patient said except for random changing of one or two words.	3.7	Therapist makes a verbal response of two or three words, given as simple acceptance and understanding of what the patient says.
4.3	Therapist states a question or incomplete sentence which contains a key word or phrase from the patient's previous response.	4.4	Therapist asks the patient to tell him more, to elaborate a little, on a topic already mentioned.
5.8	Therapist asks for "an example" of what the patient has just reported.	5.7	Therapist parries a question put to him by the patient, by directing it back to the patient.
6.2	Therapist focuses upon an objective, factual aspect of patient's life (age, job, salary).	6.3	Therapist inquires how long patient's symptoms have been present.
6.7	Therapist restates things the patient has said, in a different way, to make the import clearer.	6.7	Therapist asks when some event (described by the patient) actually happened.
7.4	Therapist asks patient how he feels about something or some event which the patient has just talked about.	7.5	Therapist question focuses upon patient's transient thoughts within the interview situation, at a particular instant.
7.8	Therapist confronts the patient with a reformulation of things the patient has said, and asks if that is what he means.	7.8	Therapist reflects a feeling or need clearly implied in the patient response, but not actually verbalized by the patient.
8.5	Therapist conveys his impression that there is something missing from the patient's story.	8.7	Therapist summarizes a number of different responses made by the patient, which are essentially concerned with the same feeling, of which the patient is aware, and therapist labels the feeling.
9.1	Therapist suggests that what the patient has just said is inconsistent with certain other things said earlier by the patient.	9.3	Therapist points out some reality condition which is inconsistent or incompatible with the patient's wishes or expectations

therapist operations. The findings of this study illustrate the representativeness of the 50 therapist responses, and a meaningful order of conventional categories along the assumed dimension of Activity. They accord, moreover, with the scheme for analysis of Activity Levels and with the principles of focus and of minimal activity set forth many years ago by Finesinger (1948).

Study 3. One of the ultimate *applied* goals of the research program was to study therapist verbal be-

havior in the initial, rather than in the treatment interview. Since initial interviews tend usually not to involve the more active types of therapist operations (e.g., interpretive), 25 of the most active responses were removed from the original set of 50. To the remaining 25 responses, 11 more were added. The new set of 36 responses was rated, as before, along an 11-point scale of Activity, by three groups of subjects. One group consisted of 15 of the original 30 psychiatrists used in Study 1. Two other groups

TABLE 4
MEAN ACTIVITY LEVEL AND STANDARD DEVIATION
FOR FIVE INITIAL INTERVIEWS (STUDY 4)

Theoretical orientation	Number of therapist responses ^a	Mean Activity Level per response over entire interview	SD
Rogers	48	6.2	1.15
Deutsch	58	5.9	1.33
Wolberg	95	5.8	1.68
Gill	150	5.2	2.04
Finesingerian	62	4.4	1.37

^a A maximum of two initial or terminal responses (e.g., a greeting) in each interview was not rated.

were drawn from a class of 100 freshmen medical students. One group of 19 subjects had previously expressed very high interest in ultimate specialization in psychiatry, while the other group of 18 subjects had expressed very low such interest. Inclusion of medical students with high and low interest provided a check upon the independence of Activity ratings from psychiatric experience and sophistication.

Results of Study 3. There was an overall between-group difference in mean Activity Level assigned the 36 responses (see Table 1). The value of *F* was 5.82 ($p < .01$), the medical students with low interest in psychiatry being most deviant from the psychiatrists. The rank orders of the responses rated by the three groups nevertheless agreed fairly well. Values of ρ ($p < .001$) are shown in Table 2. The interclass r was .49 for psychiatrists and .42 for each group of students; the reliability of average ratings was .93 for all three groups (Guilford, 1954). These values are almost identical with those found in Study 1. Indeed, the value of ρ for the mean Activity Levels of the 25 responses common to Studies 1 and 3 was .945 ($p < .001$). The data indicate that reliability of the rating procedure is but little altered by psychiatric interest and experience.

Consequently, data from the psychiatrist subjects in Study 3 were used to form two parallel Activity Scales. These are presented in Table 3. Each ordinal pair of items was matched on the basis of virtually identical mean Activity Levels and of nonsignificantly different variances.

Study 4. This study was performed to make a preliminary test of the reliability and discriminatory capacity of Scale A. The authors independently rated, in context, each therapist response in five unfamiliar published initial interviews. These, chosen for their divergence of theoretical adherence, were performed by Wolberg (1954, pp. 690-699); Deutsch (Deutsch

& Murphy, 1955, pp. 29-49); Skinner (Finesinger & Powdermaker, undated); Gill (Gill, et al., 1954, pp. 134-204), and Rogers (1947, pp. 128-142). Since the Finesinger and Powdermaker interview was actually performed by a close adherent to the Finesinger technique, all subsequent reference to this interview will be via the term "Finesingerian."

Results of Study 4. Reliability of scoring the five interviews was .90 or better. Table 4 shows mean and σ of each interview for one of the two raters. The mean values differ from each other both by Fisher's *F* and by Kruskal-Wallis' (1952) *H* ($p < .001$). This result supported the assumption that the Activity Scale samples a meaningful common variable in therapist verbal behavior, and hence justified a more powerful and elaborate reliability study.

Study 5. This was undertaken (a) systematically to assess the range of reliability estimates obtained when professional but untrained raters apply the Activity Scales to unfamiliar printed interview material; and (b) to study the empirical equivalence (i.e., the interchangeability) of the two parallel Activity Scales (see Scales A and B, Table 3).

Eight raters consisting of four clinical psychologists at the PhD level and four psychiatrists having between 2 and 4 years of experience were used in a modified latin square study adapted from Cutler, Bordin, Williams, and Rigler (1958). For each of four interviews the subject rated successive therapist responses seriatim for Activity Level, using either Scale A or Scale B. Each scale was used with a different pair of the four interviews presented to each subject. In order to control for the possibility that ratings of the therapist responses might be influenced by the succeeding response from the patient, only the therapist responses were presented for two of the interviews rated by each subject (the "Context Absent" condition); whereas for the other two inter-

TABLE 5
EXPERIMENTAL DESIGN OF STUDY 5

Rater	Scale A		Scale B	
	Context Absent ^a	Context Present ^a	Context Absent	Context Present
1. PhD	IV ^b :1 ^c	III:3	II:4	I:2
2. MD	IV:1	III:3	II:4	I:2
3. PhD	III:4	IV:2	I:1	II:3
4. MD	III:4	IV:2	I:1	II:3
5. PhD	II:2	I:4	IV:3	III:1
6. MD	II:2	I:4	IV:3	III:1
7. PhD	I:3	II:1	III:2	IV:4
8. MD	I:3	II:1	III:2	IV:4

^a "Context Absent" implies that the patient's responses were not presented to the subject; "Context Present" implies that they were so presented.
^b Roman numerals refer to a particular interview (see text).
^c Arabic numerals refer to the order of presentation to the subject, of a specific interview and a particular treatment combination.

TABLE 6
ANALYSIS OF VARIANCE

Source of variance	SS	df	MS	F
Sequences	.53	3	.177	<1
Raters within sequences	.84	4	.210	<1
Total between raters	1.37	7		
Order	1.01	3	.337	<1
Interviews	17.10	3	5.700	9.97***
Experimental conditions	3.89	3	1.297	2.20
Context	.28	1	.280	<1
Scales	3.30	1	3.300	5.77*
Context \times Scales	.31	1	.310	<1
Pooled error	8.58	15	.572	
Total within raters	30.58	24		
Grand total	31.95	31		

* $p < .05$.
*** $p < .001$.

views the entire typed protocol was presented (the "Context Present" condition). The Sequence of treatment combinations presented to each subject, the Order in which a given interview was rated, the Context variable, and the Scale variable were systematically varied and controlled in the design shown in Table 5. It will be noted that four pairs of subjects (one psychiatrist and one clinical psychologist) were treated identically with one of four sets of treatment combinations.

The four interviews were selected from those used in Study 4. They are hereafter referred to as Interviews I (Wolberg) consisting of 73 therapist responses;² II (Gill), 73 responses; III (Finesinger-

² It was necessary to ignore certain types of therapist responses (two from each interview) such as an initial greeting or a farewell.

TABLE 7
OVERALL MEAN ACTIVITY LEVEL AND SIGMA ASSIGNED
EACH OF FOUR INTERVIEWS BY EIGHT RATERS IN
STUDY 5

No.	Interview	Mean AL ^a	N ^b	σ
I	Wolberg	6.1	73	1.45**
II	Gill	5.1	73	1.71**
III	Finesingerian	4.9	62	1.10**
IV	Rogers	6.7	48	.71**

^a Based upon the pooled ratings of eight subjects.

^b Number of therapist responses rated.

** The variances are all significantly different from each other ($p < .01$).

TABLE 8
INTRRATER RELIABILITIES AS A FUNCTION OF PROFESSIONAL SPECIALTY

No.	Interview	Among PhD's		Among MD's		Among all Subjects	
		Range	Median ^a	Range	Median ^a	Range	Median ^a
I	Wolberg	.66-.76	.73	.46-.74	.56	.46-.79	.62
II	Gill	.72-.84	.78	.40-.74	.50	.40-.85	.73
III	Finesingerian	-.09-.60	.34	.26-.55	.34	-.14-.65	.35
IV	Rogers	.00-.64	.38	.13-.64	.36	-.05-.83	.39

Note.—Values of N are 73 for Interviews I and II; 62 for III; and 48 for Interview IV. The minimum r for which $p < .05$ is .23 for Interviews I and II; .25 for III; and .29 for IV. The minimum r for which $p < .01$ is .30 for Interviews I and II; .32 for III; and .37 for IV.

The median reliabilities of Interviews I and II are higher, both in the PhD group and in the "All subjects" group ($p < .01$), than those of Interviews III and IV.

^a Each median value is derived from a population of 6 r 's for the two "specialty" groups, and from 1 of 28 r 's for the "all subjects" group.

TABLE 9
 INTERRATER RELIABILITIES AS A FUNCTION OF EXPERIMENTAL CONDITIONS

Combination of experimental conditions	Interview number							
	I		II		III		IV	
	Range	Median ^a	Range	Median ^a	Range	Median ^a	Range	Median ^a
Same scale, same context	.58-.79	.69**	.41-.85	.76	.36-.64	.41	.01-.68	.35
Different scale, different context	.54-.74	.60	.40-.87	.69	-.14-.63	.37	-.02-.64	.35
Same context, different scale	.46-.74	.64	.48-.84	.71	.09-.59	.34	-.05-.83	.37
Same scale, different context	.49-.76	.73	.42-.79	.74	.19-.65	.34	.15-.64	.44

Note.—"Same context" implies that the two subjects from whose data a given r is derived, both rated *either* with patient context present, or with patient context absent; "different context" implies that one member of such a pair of subjects rated with patient context present, the other member with patient context absent. "Same scale" implies that such a pair of subjects both rated *either* with Scale A, or with Scale B; "different scale" implies that one member rated with Scale A, the other with Scale B.

For values of N , and for minimal values of r achieving significance, see the general footnote to Table 8.

^a The median r is drawn from a population of four r 's in the "same scale and same context" condition, and from one of eight r 's in each of the other three conditions.

** The median values for Interviews I and III, line 2, are significantly different ($p < .01$). See text for comments on other comparisons.

ian), 62 responses; and IV (Rogers), 48 responses. In order to keep the subject's task within a reasonable time limit, only the first 73 responses of interviews I and II were used, while the other two interviews were used in their entirety. The total time taken for the four tasks varied from 1 to 3 hours per subject.

Results of Study 5.³ The analysis of variance is shown in Table 6. The most clear-cut and crucial fact established by the analysis is that the mean Activity Levels of the four interviews significantly differ among themselves ($p < .001$). These mean values are presented in Table 7. The rank order of the four interviews accords exactly with that found in Study 4. It is noteworthy that the Rogerian interview (IV) turns out to be the most active and the Finesingerian one the least active. The sole other significant effect is for the Scale variable ($p < .05$). This accords with an a priori hunch, but it is nevertheless potentially somewhat disturbing; for it will furthermore be seen later that for three of the interviews, Scale A manifests greater reliability than does Scale B. While the context variable turns out not to be significant (which is as it should be), it should be noted, for the present, that this finding refers only to overall mean values for each interview.

The between-rater reliabilities for each interview were computed by IBM, yielding a total of $4(8 \times 7)/2 = 112$ reliability coefficients (Pearson r 's). One summary of these, presented in Table 8, breaks the data down into a PhD group (clinical psychologists) and

an MD group (psychiatrists). The PhD group shows nonsignificantly greater median within-group agreement⁴ than does the MD group on Interviews I and II. The same two interviews were rated more reliably than Interviews III and IV, by both the PhD group and the "all subjects" group ($p \leq .01$). That the Rogers interview (IV) elicited low rater agreement was not at all surprising, since many Rogerian responses do seem extremely difficult to match with scale items, and considerable argument was voiced by several subjects that their ratings of this interview were subjectively most unreliable. The equally low reliability of the Finesingerian interview (III), however, was rather surprising, and no satisfactory explanation of this finding is forthcoming.

Table 9 presents the median and range of interrater r 's as a function of the various experimental conditions. Generally speaking, the highest reliability coefficients are obtained when comparisons are made with the same scale, and lowest when they are made with ratings based upon different scales; but the differences do not achieve conventional significance. The reliabilities of both Interviews I and II are consistently arithmetically larger than those of interviews III and IV within all four experimental treatment combinations. Only 1 of the 16 possible comparisons yields a significant value of p (see Table 9, level of significance footnote), while the median value of p for the set is about .13. The effect of the Context variable is slight when assessed from the data presented in Table 9. But a somewhat different picture

³ Michael S. Black, now of the University of Illinois, performed most of the tedious office computations in Study 5.

⁴ All comparisons of r 's subsequently reported in this paper were made, unless otherwise stated, with the Median test (e.g., Siegel, 1956, 111-116).

is presented in Table 10, where the reliabilities are examined within pairs of raters (one PhD and one MD) each of the two members having been treated identically. For Interview I the Context variable produces fairly consistent correlations within pairs of subjects using either Scale A or Scale B. For Interview II, however, the correlation is greater for Scale B condition when the patient context is present than when it is absent ($p < .05$). For Interview III the latter finding holds for both scales but does not achieve conventional significance.

A surprising finding for Interview IV is that when the patient context is *present* the reliabilities drop, to near zero for Scale A ($p > .05$) and by 50% for Scale B ($p = .06$)! This type of finding was reported also by Cutler, Bordin, Williams, and Rigler (1958) whose analyst-fledgling subjects agreed significantly less in ratings of Depth of Interpretation when they had patient material available to them. In the present study this finding is taken to reflect (again) the difficulties involved in rating the Rogerian material.

A comparison of all r 's involving Scale A with complementary r 's involving Scale B (Table 10) shows that for Interviews I and II, consistently greater within-pair agreement ($p \leq .02$) is obtained with Scale A. The same comparisons for Interview III fall short of significance, although they too are in a consistent direction. For Interview IV (Rogers), on the other hand, exactly the opposite outcomes are observed, of which one is significant ($p < .05$) and the other nearly so ($p = .06$).

DISCUSSION

The empirically derived Activity Scales facilitate ratings having average reliability which is moderate (.51) for untrained raters (Study 5) and very high (.91) for well-trained raters (Study 4). The Activity Scales satisfactorily discriminate among the interviews employed. In Study 5, however, differences among the reliabilities of the four interviews are considerable; the values for Interviews III (Finesingerian) and IV (Rogers) being considerably lower than the estimates for the other two. The Rogers interview in addition leads to two unexpected discrepancies requiring discussion.

A problem facing all of those who experiment with ratings of therapist verbal behavior concerns the selection of subjects. The natural, defensible tendency is to obtain the services of highly trained "experts." This, of course, raises serious questions of practical availability, since the expert not only has less time to donate to research workers, but he is also in much shorter supply than the non-expert. While in Study 5 subjects with the

TABLE 10
CORRELATIONS WITHIN EACH PAIR OF RATERS UNDER
ALL CONDITIONS OF IDENTICAL EXPERIMENTAL
TREATMENT

Number of interview	Scale A		Scale B	
	Patient Context		Patient Context	
	Present	Absent	Present	Absent
I	.78	.79	.58	.59
II	.85	.84	.67	.41
III	.64	.40	.42	.36
IV	.01	.30	.40	.68

Note.—Each r is based upon a unique combination of variables for each interview. Professional specialty (i.e., clinical psychology vs. psychiatry) is confounded with the Sequence, Order, Context, and Scale variables in all 16 indices.

For values of N , and for minimal values of r achieving significance, see the general footnote to Table 8.

Values for Scale A are larger ($p \leq .02$) than those for Scale B within Interviews I and II. Values for Scale B are larger ($p \leq .06$) than those for Scale A in Interview IV. In Interview II, Scale B, the r for Patient Context Present is larger than that for Patient Context Absent ($p \leq .05$). In Interview IV, the r for Patient Context Absent, Scale B, is larger than that for Patient Context Present ($p = .06$). The preceding values of p were obtained by using the z transformation (McNemar, 1955, p. 148).

PhD may have shown a slight, but inconsistent edge over those with the MD, the overall results indicate that the reliability of performance is very much more a function of experimental conditions than of professional specialty. A conclusion comparable in principle was reached by Cutler et al. (1958).

The two Activity Scales not only led to different overall mean ratings of Activity Level; interrater reliabilities also differed as a function of the particular scale. Scale A was more reliably employed for three of the interviews, Scale B being more reliably used with Interview IV (Rogers). This is somewhat alarming, because the selection of particular illustrative points along the empirical scale dimension was in the present case (and presumably was in several other reported studies—e.g., Harway, Dittman, Raush, Bordin, & Rigler, 1955) largely an arbitrary matter. The empirical differences between the scales thus raise an important theoretical issue which now deserves comment.

On the one hand it is quite possible that the two scales have different dimensionalities, Scale B being, say, two-dimensional, and Scale A one-dimensional. (One-dimensionality of the Activity continuum has heretofore been

assumed, but not empirically proven.) On the other hand, it is also possible that the (Rogerian) responses in Interview IV are in toto two-dimensional, while those of the other three interviews can be adequately represented with a single dimension. Granted these two contingencies then it would follow that, with Interview IV, Scale B would elicit greater interrater reliability than Scale A. It is furthermore suggested that *Interpretation* may constitute this second dimension among both the items of Scale B and the therapist responses in Interview IV.

The plausibility of the foregoing hypothetical argument may be clearer in the light of the following considerations. The "reflection," which is the basic and most frequent verbal operation in the Rogerian interview (Rogers, 1951) presumably takes one some distance along a dimension of interpretation. In contrast, the other three interviews used in Study 5 between them contain less than a half-dozen responses that could be classified as "interpretive." Furthermore, inspection of Table 3 shows that Item 8 in Scale B contains the sole reference (in either scale) to "reflection of feelings." It is likely that subjects employed this category to classify those responses in Interview IV which were typically Rogerian in nature,⁵ whereas in Scale A no comparable item lay at the subject's disposal.⁶ Consequently, the reliability of Interview IV would turn out, as suggested above, to be higher with Scale B than with Scale A.

When one speaks of Scale B as facilitating "higher reliability" of ratings for Interview IV it must be noted, however, that ratings of therapist responses in this interview markedly *drop* under both scale conditions when patient context is added. This finding is quite

⁵ At least 5 of the subjects were clearly aware that Interview IV was Rogerian.

⁶ A rough check bearing out the tenability of this hypothesis is as follows. A frequency count across all eight raters was made of the frequencies with which, for each interview, the *eighth item* in Scale A and Scale B were employed. For Interview I, the respective proportions of responses classified in Item 8 were .15 for Scale A, and .02 for Scale B. For Interview II the respective proportions were .02 and .04; and for Interview III, .00 and .10. For the Rogerian interview (IV), however, the proportions were .30 for Scale A, and .53 for Scale B.

opposite to that for Interviews I, II, and III; and furthermore is not in accordance with logical expectation. For, depending upon the nature of some given "exploratory question," say, from the therapist, there should be predictable effects upon rating-reliability, of the addition of patient context. If the referents of the therapist question are absolutely clear to the rater (i.e., if the referents are completely defined by the question per se) then addition of patient context should not affect the reliability of rated Activity Level. If, however, the referents of the therapist question are *not* entirely clear to the rater, then addition of patient material should raise (but never lower) the reliability of rated Activity Level for the particular question.

The foregoing suggests that the very presence of patient context during rating of Rogerian responses in Interview IV somehow undermined the subject's understanding of what a Rogerian reflection of feeling looks like. Indeed, at least in the particular interview studied here, the reflection frequently does not seem, subjectively, to bear any consistent contextual relation to whatever follows from the patient.

From the standpoint of theory and research it is desirable to examine in more detail this question of dimensionality with respect to both Scale B and the Rogerian therapist behavior of Interview IV, in hope that a modified scale might be assembled having high reliability with both non-Rogerian and Rogerian material. But this whole issue is of course an applied offshoot of the more general and fundamental question of the dimensional relations between *elicitation* of information, and *interpretation* of information (which according to the results of Studies 1 and 2 involves relatively high Activity Level). A relevant study along the lines of one by Raush, Sperber, Rigler, Williams, Harway, Bordin, Dittmann, and Hays (1956) is to be performed in the near future.

In summary, it is felt that, subject to the restrictions outlined above which demand further clarification, we have examined some of the critical variables affecting the reliability of ratings of therapist Activity Level; and that the scales themselves are sufficiently meaningful and reliable to justify their ap-

plication in further research on the initial as well as the therapeutic interview. Attention may now be turned toward specification and examination of relevant variables in the patient's behavior as a function of therapist Activity Level.

SUMMARY

This paper describes the development of a parallel pair of scales for assessing the Activity Level of discrete Therapist Verbal Responses, and the application of the scales to several published initial interviews. In Study 1, 30 Board-certified psychiatrists rated 50 abstract descriptions of Therapist Verbal Responses along an 11-point scale of "Activity," the latter being defined in terms of the degree of "Ambiguity, Lead, and Inference." Interjudge reliability was .50, and the intraclass r , .93. Each rater also categorized the 50 responses according to whether he considered them primarily used for purposes of treatment, or for purposes of diagnosis. Those therapist responses agreed to be primarily "therapeutic" in purpose were rated with a considerably higher mean Activity Level than others classified as "diagnostic" in purpose.

In Study 2 it was shown that a large majority of the 50 therapist responses was agreed by independent judges to typify one of the following conventional categories of therapist operation: Simple Facilitation, Exploration, Clarification, Interpretation, and Supportive Reassurance. The responses classified in these successive categories, respectively, showed increasingly higher mean Activity Levels. Consequently, it was assumed that the main set of 50 responses included representative elements from the entire range of typical therapist operations.

Study 3 involved further rating of a revised set of 36 Therapist Verbal Responses belonging *mainly* in the categories of Simple Facilitation, Exploration, and Clarification. The subjects consisted of 15 of the psychiatrists used in Study 1, and 37 freshmen medical students, 19 with high, and 18 with low interest in psychiatry. Reliability of ratings was only slightly lower for student subjects than for psychiatrist subjects; and students with low interest in psychiatry showed least (though still highly significant) agreement

with the psychiatrists' rank ordering of the therapist responses. Since the rating procedure did not appear to be a serious function of either professional interest or experience, a parallel pair of 10-point scales of Activity Level were assembled using data from the psychiatrist subjects.

In Study 4 the individual therapist responses of five unfamiliar published initial interviews were rated by both authors, using Scale A. Interjudge correlation was .90 or better. A more elaborate and rigorous reliability study was then performed.

In Study 5 the two Activity Scales were then employed in a latin square design requiring eight untrained raters (four psychiatrists and four clinical psychologists) to rate for Activity Level the therapist responses in four widely differing published initial interviews (by Wolberg, by Gill, by a Finesingerian, and by Rogers). Scale A vs. Scale B constituted one factorial variable, and Patient Context Absent vs. Present constituted the other. The analysis of variance showed a significant difference among the interviews, and a significant main effect for the Scale variable. When interjudge reliabilities were examined the two types of subjects (psychiatrists and psychologists) showed only minor differences. Further, the Wolberg and the Gill interviews were consistently more reliably rated than were the other two. Scale A, however, was consistently more reliably employed than Scale B with three of the four interviews, but Scale B was more reliably employed with the fourth (Rogerian) interview. Furthermore, while adding Patient Context either increased or did not affect reliability of rating (with either scale) of the first three interviews, the reliability of rating the Rogerian interview clearly *decreased*.

The discrepancies involving the Rogerian interview were discussed, and a hypothetical basis for their occurrence was advanced which concerned the dimensionalities of the two scales and of Rogerian vs. non-Rogerian therapist responses. It is concluded that while the general problem of dimensionality needs further examination, we have a pair of parallel Activity Scales the reliabilities of which are comparatively satisfactory (the grand median of 112 coefficients is .50), and

that we have explored some of the conditions likely to affect their application by untrained, professional raters. One may now turn toward investigation of patient variables as a function of Therapist Activity Level.

REFERENCES

- BERNSTEIN, A., LENNARD, H. L., & PALMORE, E. Ease of communication during psychotherapy. Technical Report No. 5, 1958, Columbia University, Bureau of Applied Social Research, Psychotherapy Interaction Research Project.
- BORDIN, E. S. Ambiguity as a therapeutic variable. *J. consult. Psychol.*, 1955, 19, 9-15.
- CUTLER, R. L., BORDIN, E. S., WILLIAMS, J., & RIGLER, D. Psychoanalysts as expert observers of the therapy process. *J. consult. Psychol.*, 1958, 22, 335-340.
- DEUTSCH, F., & MURPHY, W. F. *The clinical interview*. Vol. 2. New York: International Univer. Press, 1955.
- DIBNER, A. S. The relationship between ambiguity and anxiety in a clinical interview. Unpublished doctoral dissertation, University of Michigan, 1953.
- FINESINGER, J. E. Psychiatric interviewing: Principles and procedure in insight therapy. *Amer. J. Psychiat.*, 1948, 105, 187-195.
- FINESINGER, J. E., & POWDERMAKER, F. A clinical picture of claustrophobia. Part V. In, *Seven films on interviewing and psychotherapy*. Washington, D. C.: Veterans Administration, Department of Medicine and Surgery, undated.
- FREUD, S. Fragment of an analysis of a case of hysteria. (Originally published 1905.) In *Collected Papers*. Vol. 2. London: Hogarth, 1948.
- GILL, M., NEWMAN, R., & REDLICH, F. C. *The initial interview in psychiatric practice*. New York: International Univer. Press, 1954.
- GUILFORD, J. P. *Psychometric methods*. New York: McGraw-Hill, 1954.
- HARWAY, N. I., DITTMAN, A. T., RAUSH, H. L., BORDIN, E. S., & RIGLER, D. The measurement of depth of interpretation. *J. consult. Psychol.*, 1955, 19, 247-253.
- KRUSKAL, W. H., & WALLIS, W. A. Use of ranks in one-criterion variance analysis. *J. Amer. Statist. Ass.*, 1952, 47, 583-621.
- LINDQUIST, E. F. *Design and analysis of experiments in psychology and education*. Boston: Houghton Mifflin, 1953.
- MCMENAR, Q. *Psychological statistics*. New York: Wiley, 1955.
- MANN, H. B., & WHITNEY, D. R. On a test of whether one of two variables is stochastically larger than the other. *Ann. math. Statist.*, 1947, 18, 50-60.
- OSBURN, R. G. An investigation of the ambiguity dimension. Unpublished doctoral dissertation, University of Michigan, 1951.
- RAUSH, H. L., SPERBER, Z., RIGLER, D., WILLIAMS, JOAN V., HARWAY, N. I., BORDIN, E. S., DITTMAN, A. T., & HAYS, W. A dimensional analysis of depth of interpretation. *J. consult. Psychol.*, 1956, 20, 43-48.
- ROGERS, C. R. The case of Mary Jane Tilden. In W. U. Snyder (Ed.), *Casebook of non-directive counseling*. Cambridge: Riverside, 1947.
- ROGERS, C. R. *Client-centered therapy*. Boston: Houghton Mifflin, 1951.
- SIEGEL, S. *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill, 1956.
- SNYDER, W. U. An investigation of the nature of nondirective psychotherapy. *J. gen. Psychol.*, 1945, 33, 193-223.
- WOLBERG, L. R. *The technique of psychotherapy*. New York: Grune & Stratton, 1954.

(Received September 26, 1960)

THE ACCURACY OF CLINICAL PSYCHOLOGISTS' ESTIMATES OF INTERVIEWEES' INTELLIGENCE¹

ZANWIL SPERBER AND ARTHUR M. ADLERSTEIN

Children's Hospital of Philadelphia, Pennsylvania

That accurate appraisal of intellectual capacity can best be accomplished using standardized test procedures is well accepted. There are often occasions, however, when clinical decisions are influenced by judgments of intelligence which must be based on observations rather than tests. It is therefore important to ask, "How good are clinicians' estimates of intelligence?" The purpose of this paper is to present data indicating the relationship between clinical psychologists' estimates of intelligence, based on observations of only the verbal behavior of interviewees, and psychometric measures of the interviewees' intelligence.

METHOD

Subjects. Five clinical psychologists served as judges. Four were PhDs. All had substantial experience with intelligence testing of adults and children, and were familiar with current approaches to the conceptualization and measurement of intelligence.

Interviewees. The women whose IQs were estimated had been interviewed as part of a follow-up study of their children, 4-6 years of age, who had

been treated medically at birth (see Footnote 1). The sample size was set at 30 because the time required for data collection given this *N* reached the limit volunteer judges could be asked to contribute. We also felt this *N* was large enough to yield meaningful statistical results.

The 30 cases were drawn using a stratified random procedure to be representative of the range and mean of the intelligence test scores of the larger follow-up group. A social class categorization of the husbands' occupations (Sarason & Mandler, 1952; Sperber, 1959) shows that the sample included members of the middle, lower middle, skilled, and unskilled worker classes. Table 1 shows the age, education, and measured IQ of the sample.

Interviews. The hour-long interviews were conducted by psychiatrists and focused on the child's developmental progress, and on maternal attitudes. Attempts were made to elicit some description of the mother's life history. The interviews were tape recorded and verbatim transcripts typed. Judges 1, 2, and 3 made intelligence estimates after reading the transcripts, Judges 4 and 5 after hearing tape recordings. The judges had no other contact with the mothers.

Intelligence criteria. An abbreviated WAIS (Wechsler, 1955) consisting of four subtests, Vocabulary, Information, Block Design, and Picture Arrangement was routinely administered to the mothers. The four subtests give a good approximation to the IQ obtained using the full scale (Cohen, 1957; Doppelt, 1956; Himelstein, 1957). Hereafter the prorated IQ based on scores on the four subtests will be called

TABLE 1
AGE, EDUCATION, AND MEASURED IQ
OF THE INTERVIEWEE

Variable	<i>N</i> ^a	Mean	<i>SD</i>	Range
Age (in years)	30	36.2	6.5	25-54
Years of education	25	11.8	1.9	8-16
IQ criteria				
WAIS ^b	29	101.7	13.5	68-135
Vocabulary	30	104.5	18.1	70-149

^a Number of cases with relevant data available.

^b Based on four subtests: Vocabulary, Information, Block Design, Picture Arrangement.

¹ The interviews and intelligence test data used in the present research were collected as part of a study of children who had had blood problems as neonates, in most cases involving Rh incompatibility and treated by exchange transfusions. The follow-up study was supported by the National Institute of Neurological Diseases and Blindness, National Institutes of Health, United States Public Health Service as part of the Collaborative Project to Study the Etiology of Cerebral Palsy and Other Neurological Diseases of Infancy and Childhood.

T. McNair Scott is Senior Investigator for the Collaborative Project at Children's Hospital, and T. R. Boggs, Jr., pediatrician; C. Kennedy, neurologist; and J. A. Rose, psychiatrist, are co-investigators for the Collaborative Project and the follow-up study.

We appreciate the contribution made by our psychologist colleagues who served as judges, and Elizabeth Hirshman's assistance with the statistical computations.

TABLE 2
CORRELATIONS BETWEEN JUDGES' IQ ESTIMATES
AND OTHER CRITERIA

IQ criteria	N	Judge					Mean r
		1	2	3	4	5	
Wais	29	.43*	.82	.76	.72	.69	.70
Vocabulary	30	.67	.80	.78	.72	.76	.75
"Other Judges" ^a	30	.74	.77	.72	.77	.78	.76

^a Entries are mean correlations between indicated judge's estimates and those of the other four judges.
* < .05; all other correlations are statistically significant at the $p < .01$ level.

the WAIS IQ. WAIS IQs were available for 29 women. In the last phase of the follow-up study only the Vocabulary subtest was given. One of the mothers inadvertently included in our sample was from the later group.

Since verbal production served as the judges' primary source of information about the interviewees' intelligence, a prorated IQ based only on the interviewees' scores on the WAIS Vocabulary test was used as a second criterion. Hereafter the prorated IQ based on the Vocabulary subtest will be called the Vocabulary IQ.

Procedure. Judges were asked to assess the interviewees' IQs with no further discussion of how they should define intelligence or use the interview material. They made their judgments independently, specifying a exact number for estimates between 70 and 140, and indicating after each estimate whether the judgment had been made with high or low confidence.

Judges were aware of the general nature of the follow-up study and knew the mothers had taken an abbreviated WAIS. For each case they were told the age of the child who was the subject of the interview. Cases were prearranged to form five random sequences of IQs. Each judge followed a different sequence in making his estimates.

RESULTS

Correlational analyses. Product-moment correlations were calculated between each judge's estimates and (a) the WAIS IQ, (b) the Vocabulary IQ, (c) each other judge's IQ estimates. The 10 coefficients between estimated and measured IQ, presented in Table 2, are positive and, with one exception, substantial. The estimates of all possible pairs of judges were positively correlated at the .01 level. Table 2 shows the mean correlations between each judge's estimates and those of the other judges. Presence of voice cues did not influence the correlations between judges' estimates, or between estimates and criteria.

Discrepancies between measured IQ and estimated IQ. The mean IQ assigned by each of the five judges ranged from 100.5 to 105.3, corresponding closely to the test means shown in Table 1. The SDs of judges' estimates were somewhat smaller than the SDs of test scores. Four judges restricted their estimates to an 80-126 IQ range.

As indicated in Table 3 judges' estimates often deviated appreciably from the measured IQs. Over all five judges the mean discrepancy between the estimated IQs and WAIS IQs was 7.8 points. The mean discrepancy from the Vocabulary IQ was 9.9 points. Considering only the WAIS criterion in relation to which judges' estimates were more accurate, 83% of the estimates made by the most accurate judge, and 66% made by the least accurate judge, were within 10 IQ points of the criterion. Over all five judges, 72% of the estimates were within this range.

Table 3 presents the result of two additional analyses. Knowing that the WAIS was standardized so that the mean IQ score of a sample representative of the "population at large" would be 100 (Wechsler, 1955) how accurate would a clinician be if he simply "programed" himself to estimate each inter-

TABLE 3
DISCREPANCIES BETWEEN ESTIMATED IQ
AND MEASURED IQ

Source of estimate	IQ criteria ^a	Discrepancy		
		Mean	SD	Range
Judge				
1	WAIS	8.9	6.4	0-25
	Vocabulary	10.3	9.1	1-39
2	WAIS	5.4	4.7	0-17
	Vocabulary	9.2	7.9	1-31
3	WAIS	7.0	5.5	1-18
	Vocabulary	9.8	7.2	1-25
4	WAIS	9.7	7.5	0-28
	Vocabulary	10.4	7.8	1-35
5	WAIS	8.2	6.3	1-25
	Vocabulary	9.6	7.2	1-28
Assumed population mean (IQ = 100)	WAIS	11.0	8.0	0-35
	Vocabulary	15.1	11.0	0-49
WAIS IQ vs. Vocabulary IQ		6.9	6.3	0-21

^a N = 29 for WAIS IQ criterion; N = 30 for Vocabulary criterion.

viewee's IQ as 100? ² As shown in Table 3, the mean, *SD*, and range of discrepancies obtained by a hypothetical programed judge would have been larger than those of our judges. Table 3 also shows the mean, *SD*, and range of discrepancies between the interviewees' WAIS IQs and their Vocabulary IQs. Despite test overlap, the discrepancies between the two psychometrically derived IQs are not appreciably less than those observed between judges' estimates and the WAIS criterion.

Judges' confidence and their IQ estimates. The five judges differed markedly with respect to the confidence with which they made the IQ estimates. Judge 1 was highly confident on only four estimates, Judges 4, 5, and 3 on 13, 15, and 16 estimates, respectively, while Judge 2 was highly confident on 23 estimates. This degree of variability suggests that the source of confidence was unique to the judge and not a function of some observable attribute of the interview material. To test this supposition judges were paired in all 10 possible combinations. The percentage of cases where both judges agreed in feeling either high or low confidence was compared to the percentage of agreements expected by chance. Agreement ranged from 30% to 53%, falling below chance expectancy for five pairs of judges.

The degree to which judges tended to feel highly confident after making IQ estimates was examined in relation to their performance. The judges were ranked for confidence level, assigning Rank 1 to the judge with the largest number of high confidence estimates. The criterion for a judge's performance as an IQ estimator was the average discrepancy between his IQ estimates and the WAIS IQs, the judge with the smallest average discrepancy being assigned Rank 1. The obtained rank-order correlation was .90, significant at $p < .05$ (Senders, 1958, p. 545, Table M).

Within judges there was a slight but consistent reversal in the relationship between confidence and accuracy. Biserial correlations between the judge's confidence (high or low) and size of the discrepancies between his estimates and the WAIS IQs ranged from .03 to

.18, indicating that judges tended to be more accurate on those estimates they made with less confidence.

DISCUSSION

The results are consistent with the findings of other investigators who have reported correlations between intelligence estimates made without benefit of psychometric data and intelligence test scores (Hanna, 1950; Marsh & Perrin, 1925; Wilson, 1954). Substantial discrepancies between measured and estimated IQs did occur in individual cases, but 72% of the judges' estimates were within 10 points of the WAIS criterion, and those of the more accurate judges did not deviate from the WAIS IQ any more than did a measured IQ based on the WAIS Vocabulary score.

Judges apparently are realistic in deciding how much confidence, in general, to place in their IQ estimates. There was a direct relationship between the number of judgments made with high confidence and the accuracy of judges. Written comments volunteered by three judges suggest that the contrary tendency for all judges to be a little more accurate on low confidence judgments compared to their *own* high confidence judgments was a function of their considering additional aspects of the interviews on cases perceived as difficult to evaluate.

Some of the larger discrepancies in judgment occurred because judges overestimated the IQ of the cases of low intelligence and underestimated the high extremes. A trained observer's estimate is, therefore, not to be considered a substitute for a good intelligence test where precise data are required, although it may be sufficient when a general idea of the client's intellectual capacity is all that is needed. Within these limits the present study indicates that experienced psychologists can make clinically useful estimates of interviewees' intelligence. The findings should not be generalized to teachers, parents, physicians, or other judge groups without further research.

SUMMARY

Five clinical psychologists estimated the IQs of 30 mothers who had been interviewed by psychiatrists, three judges after reading

² We wish to thank our colleague, Edna Small, who suggested this analysis.

transcripts and two after hearing tape recordings of the interviews. IQ estimates were compared with the prorated IQ based on WAIS subtests.

Correlations between estimates and the WAIS criterion were significant (mean $r = .70$), with 72% of the estimates within 10 points of the criterion. More confident judges were more accurate in their estimates.

The results indicate that experienced clinical psychologists can make useful estimates of interviewees' intelligence.

REFERENCES

- COHEN, J. A factor-analytically based rationale for the Wechsler Adult Intelligence Scale. *J. consult. Psychol.*, 1957, 21, 451-457.
- DOFFELT, J. E. Estimating the full scale score on the Wechsler Adult Intelligence Scale from scores on four subtests. *J. consult. Psychol.*, 1956, 20, 63-66.
- HANNA, J. V. Estimating intelligence by interview. *Educ. psychol. Measmt.*, 1950, 10, 420-430.
- HIMELSTEIN, P. A comparison of two methods of estimating full scale IQ from an abbreviated WAIS. *J. consult. Psychol.*, 1957, 21, 246.
- MARSH, SARAH E., & PERRIN, F. A. C. An experimental study of the rating scale technique. *J. abnorm. soc. Psychol.*, 1925, 19, 383-399.
- SARASON, S. B., & MANDLER, G. Some correlates of test anxiety. *J. abnorm. soc. Psychol.*, 1952, 47, 810-817.
- SENDERS, VIRGINIA L. *Measurement and statistics*. New York: Oxford Univer. Press, 1958.
- SPEERBER, Z. The test anxiety questionnaire: Scoring norms for a noncollege population. *J. abnorm. soc. Psychol.*, 1959, 58, 129-131.
- WECHSLER, D. *Manual for the Wechsler Adult Intelligence Scale*. New York: Psychological Corporation, 1955.
- WILSON, J. W. Correlation of clinical estimates with test scores on mental ability and personality tests. *J. clin. Psychol.*, 1954, 10, 97-99.

(Received October 3, 1960)

STIMULUS GENERALIZATION IN BRAIN DAMAGED CHILDREN¹

SARNOFF A. MEDNICK

University of Michigan

AND

CYNTHIA WILD

Yale University

Stimulus generalization (SG) can be said to have occurred when a response previously trained to be elicited by stimulus O can also be elicited by test stimuli similar to O. This phenomenon has been used extensively in explanation of verbal learning (Gibson, 1940), social activity (Hull, 1950), and clinical behavior (Mednick, 1958b). A study by Mednick has suggested that the behavioral deficit of the brain damaged adult usually described by the term "concrete" may also be understood in terms of SG. He found that the SG responsiveness of these patients was sharply curtailed (Mednick, 1955). Research has suggested that the concreteness observed in the brain damaged adult has its counterpart in the child. For example, Cotton was particularly struck by the similarity between her group of children suffering from cerebral palsy and the brain damaged adult (Cotton, 1941). In view of these findings, it seemed advisable to compare the generalization reactivity of brain damaged and intact children.

In terms of the previously obtained results with adults, it was hypothesized that brain damaged children would demonstrate less SG than intact children.

METHOD

Apparatus. The apparatus was adapted from one devised by Brown, Bilodeau, and Baron (1951). It consisted of a horizontal row of 11 lamps fastened to a flat black, curved plywood panel 6 feet by 2 feet, mounted on its long edge on a table. The lamps were spaced 9 degrees apart and were equidistant from the subject's eyes when the subject was seated

directly in front of, and 3.5 feet away from, the center lamp. The lamps will be designated by Numbers 1 through 11, with Lamp 1 being on the left of the subject, Lamp 11 on the right of the subject, and Lamp 6 being the center lamp. (The lamps used in this study were 1, 3, 5, 6, 7, 9, 11.) A red jeweled flashlight lamp, 2 inches above the center lamp, served as a fixation point and a ready signal. The reaction key was placed on the subject's lap. The experimenter was seated behind the panel out of the subject's view.

Subjects. Thirty-six children were tested in this study. Eighteen were patients at the Cerebral Palsy Clinic of the Children's Hospital in Boston. The majority of them were diagnosed as spastic, but several athetoid cases were also included. There were 10 boys and 8 girls, ranging in age from 6-6 to 14-10 years, and in intelligence from Mental Defective to Superior. The 18 children in the Control group (16 boys and 2 girls) were equated with the Cerebral Palsy (CP) group for age and IQ. Five mental defective children, with no evidence of organic brain damage, were matched individual for individual with respect to IQ and age with the 5 cerebral palsied children with subnormal IQs; the remaining 13 subjects of normal intelligence in the Control group were taken from a previous study (Mednick & Lehtinen, 1957). None of the CP children used in the study had a known visual defect. They all had full use of at least one arm and hand.

Procedure. Subjects were set to lift their hand from the reaction key as quickly as possible when the center lamp was lit. They were told that other lamps would be lit occasionally, but that they were only to respond to the center lamp. Subjects were encouraged to respond as quickly as possible. The latency of response was measured to the nearest one-hundredth of a second with a Standard Electric Timer. Two criteria were decided upon to determine whether the subject was capable of performing the task. First, the experimenter went through the instructions with the subject as many times as was necessary for him to be able to repeat them correctly. Somewhat more explanation usually proved necessary for the CP child than for the intact child. Secondly, a behavioral test of the subject's ability to understand and perform the task was also employed. After the instructions, the subject received two demonstration-test trials. If the subject responded inappropriately, he was discarded. No in-

¹ The authors wish to express their appreciation for the help and advice of Edith M. Taylor of Children's Hospital, Boston, Massachusetts, and Chipman of the Fernald School, Waltham, Massachusetts. The work was partially supported by a United States Public Health Service Grant No. M 1519 to the senior author while he was at Harvard University.

TABLE 1

PROPORTION OF SUBJECTS THAT RESPONDED AT LEAST ONCE AND TOTAL NUMBER OF RESPONSES AT EACH TEST LAMP

Group	Test lamp						
	1	3	5	6	7	9	11
Proportion of group responding at least once							
Cerebral palsy	.33	.16	.50	1.00	.39	.45	.33
Control	.50	.39	.78	1.00	.72	.67	.56
Total number of responses							
Cerebral palsy	6	3	9		9	8	6
Control	11	8	19		20	12	11

tact children were discarded; 10 CP children were discarded.

Ten consecutive training trials with the center lamp (10-15 seconds intertrial intervals) were then given. The training trials were followed without warning by a test series during which six of the peripheral lamps (Lamps 1, 3, 5, 7, 9, 11) were presented twice each, interspersed with 17 "booster" trials with the center lamp in a counterbalanced order. The total number of trials in the test series was 29, 17 with the center lamp and 12 with the peripheral lamps. Zero, one, two, or three center lamp booster trials intervened between successive test trials with the peripheral lamps. Six different orders were used for the test trials, each order beginning with a different peripheral lamp. Three subjects were assigned to each order from the CP and Control groups. Approximately 50% verbal rein-

forcement was used to keep the subject concentrated on the task and to promote optimal reaction times.

RESULTS

In previous research using voluntary response measures of SG responsiveness, no relationship has been found to exist between latency and frequency measures of SG (Gibson, 1939; Mednick, 1955; Mednick & Freedman, 1960; Rosenbaum, 1953) except under special conditions (Mednick, 1958a). These results were also observed in this experiment. The two groups did not differ significantly in mean latency of response on the training trials (the mean latency for the CP group was .393; the mean latency for the Control group was .334) nor was there a relationship between latency and frequency of response when these variables were dichotomized and subjected to chi square analysis.

The frequency generalization data are presented in Table 1 in the form of the proportion of subjects in each group responding at least once to a given lamp. As can be seen, the CP group showed less SG responsiveness than the Control group at every lamp. This is also reflected in a count of the total number of responses made at each lamp by the two groups (also in Table 1).

The first SG test trial is considered an important indicator of SG responsiveness, since it is relatively untainted by the effects of discrimination and extinction. On this test trial 14 of the 18 Control subjects responded,

TABLE 2

FREQUENCY DISTRIBUTION COMPARING GROUPS ON STIMULUS GENERALIZATION RESPONSIVENESS

Number of SG responses	Number of individuals	
	Cerebral palsy	Control
0	3	1
1	1	1
2	6	2
3	4	1
4	4	2
5		6
6		1
7		2
8		2
Total	18	18

while only 7 of the 18 CP subjects responded. This difference is significant (chi square, corrected = 4.11, $df = 1$, $p < .05$).

Table 2 presents a frequency distribution comparing the SG responsiveness of the CP and Control children. While none of the CP children gave more than four SG responses, 11 or 61% of the Controls showed five or more responses. The group differences are significant (chi square = 15.84, $df = 2$, $p < .01$). This test was performed by combining Rows 0-2, Rows 3 and 4, and Rows 5-8, collapsing Table 2 into a 3×2 table.

DISCUSSION

The hypothesis that the brain damaged children would evidence a diminished degree of SG responsiveness is supported by the results. It seems likely that this finding may help explain the behavior of the brain damaged child, which has been described as concrete. An often-cited clinical example of concrete behavior concerns the child who has been trained to complete a task seated in a certain way at a certain table. However, when his position is altered or table changed, he is no longer able to perform the task. Clearly, this could also be explained as an instance of failure of SG. The second stimulus situation differed from the first; SG did not occur.

This way of thinking of the problems of these children has certain advantages. For one thing, we can look at the teaching materials for these children in a more differentiated manner. If we want the child to respond with the same response to two different stimulus situations (grasping "abstract" concept), we should eliminate all unessential differences in the stimuli, since these will hamper generalization. In addition, we have an experimental literature in SG (recently reviewed by Mednick & Freedman, 1960), on which we can draw for suggestions or manners to augment SG responsiveness. Thus, it has been shown that greater SG responsiveness is manifested under higher drive levels (Brown, 1942; Mednick, 1957; Rosenbaum, 1953). In addition, within limits, greater training in giving a response to a stimulus will result in augmented SG responsiveness to similar stimuli (Margolius, 1955; Thompson, 1959).

SUMMARY

The hypothesis that brain damaged children suffer reduced SG responsiveness was tested and supported. SG was measured along a visual-spatial dimension with an apparatus that required a voluntary response. Some observations were made regarding the implications of this finding for the training of the brain damaged child.

REFERENCES

- BROWN, J. S. The generalization of approach responses as a function of stimulus intensity and strength of motivation. *J. comp. Psychol.*, 1942, 33, 209-226.
- BROWN, J. S., BILODEAU, E. A., & BARON, M. R. Bidirectional gradients in the strength of a voluntary response to stimuli on a visual-spatial dimension. *J. exp. Psychol.*, 1951, 41, 52-61.
- COTTON, C. B. A study of the reactions of spastic children to certain test situations. *J. genet. Psychol.*, 1941, 58, 27-44.
- GIBSON, E. J. Sensory generalization with voluntary reactions. *J. exp. Psychol.*, 1939, 24, 237-253.
- GIBSON, E. J. A systematic application of the concepts of generalization and differentiation to verbal learning. *Psychol. Rev.*, 1940, 47, 196-229.
- HULL, C. L. A primary social science law. *Scient. Mon.*, 1950, 71, 221-228.
- MARGOLIUS, G. Stimulus generalization of an instrumental response as a function of the number of reinforced trials. *J. exp. Psychol.*, 1955, 49, 105-111.
- MEDNICK, S. A. Distortions in the gradients of stimulus generalization related to cortical brain damage and schizophrenia. *J. abnorm. soc. Psychol.*, 1955, 51, 536-542.
- MEDNICK, S. A. Generalization as a function of manifest anxiety and adaptation to psychological experiments. *J. consult. Psychol.*, 1957, 21, 491-494.
- MEDNICK, S. A. Latency generalization gradients of a voluntary response. *Amer. J. Psychol.*, 1958, 71, 752-755. (a)
- MEDNICK, S. A. A learning theory approach to research in schizophrenia. *Psychol. Bull.*, 1958, 55, 316-327. (b)
- MEDNICK, S. A., & FREEDMAN, J. L. Stimulus generalization. *Psychol. Bull.*, 1960, 57, 169-200.
- MEDNICK, S. A., & LEHTINEN, L. E. Stimulus generalization as a function of age in children. *J. exp. Psychol.*, 1957, 53, 180-183.
- ROSENBAUM, G. Stimulus generalization as a function of experimentally induced anxiety. *J. exp. Psychol.*, 1953, 45, 35-43.
- THOMPSON, R. F. Effect of acquisition level upon the magnitude of stimulus generalization across sensory modality. *J. comp. physiol. Psychol.*, 1959, 52, 183-185.

(Received October 5, 1960)

THE EFFECT OF INSTRUCTIONAL TIME INTERVAL AND SOCIAL DESIRABILITY ON THE VALIDITY OF A FORCED-CHOICE ANXIETY SCALE

LAWRENCE P. BERGS AND BARCLAY MARTIN

University of Wisconsin

If one conceptualizes anxiety as an emotional arousal state that varies from situation to situation and even in the same situation from time to time, it becomes important to take this variability into account in the construction of any paper and pencil test designed to measure the construct. Ordinarily the instructions accompanying most paper and pencil tests of personality traits are vague with respect to the time interval for which the person is rating himself, although the usual implication is that the person is answering the items in terms of how he has generally been during his life.

The first aim of the present research was to administer the same anxiety scale to separate groups with instructions given in terms of three different time intervals, and observe how the correlations of the scale with a subsequent criterion assessment of anxiety in a specific stress situation were affected. The time intervals selected were "last 2 weeks, last 6 months," and "in general." The criterion assessment was made in individual sessions about a month later. It was thought that the "last 2 weeks" and "in general" groups would show less correlation with the criterion measure than the "last 6 months" group. "Last 2 weeks" should be low because this time interval would be most affected by transient fluctuations, and "in general" should be low because it covers too long a time interval, whereas "last 6 months" might more likely tap the more recent characteristic level of anxiety.

The second aim of the study was to vary the degree to which variance attributable to the tendency to respond in terms of the social desirability of the item was removed from the scale. An attempt to remove this

source of variance was made by applying different scoring procedures to a forced-choice format of item triplets similar to that used by Heineman (1953). One scoring procedure was thought to minimize social desirability variance, a second to be heavily affected by it, and a third was devised to measure more directly the social desirability variance itself. It was expected that the procedure that minimized social desirability variance would yield scores most highly correlated with the criterion.

It was also expected that social desirability variance, itself, would be differentially affected by the three instructional time intervals. First, it is reasonable to suppose that a person would be more willing to admit to socially undesirable attributes if he were saying that these were true for a 2-week period than for 6 months or in general, and accordingly the mean social desirability scores should decrease with increasing time intervals. Secondly, if one of the scoring procedures does indeed successfully remove part of the social desirability variance, then the mean scores for that procedure should vary less as a function of instructional time interval than the scores for the procedure that includes more of this variance.

Previous research is meager with respect to showing differential validity as a function of the degree to which social desirability variance is removed from paper and pencil anxiety measures. Edwards (1957) has pointed up the pervasiveness of this variance and developed a scale to measure it. Heineman (1953) attempted to rid the Taylor *MA* scale of this variance by constructing a forced-choice version, and showed that the correlation with the *MMPI K* scale could be re-

duced. Silverman (1957) found that Heineman's forced-choice form correlated .24 ($p = .05$) with base level palmar conductance obtained before a stress session, whereas the regular Taylor *MA* scale correlated only $-.02$. Martin (1959) reported a correlation of .44 between base level palmar conductance during stress and a forced-choice scale composed of adjective triplets taken immediately after the stress session and a correlation of $-.02$ between the same measure and the regular Taylor *MA* scale taken earlier in a group session. The correlation of .44 was obtained in a group that took the adjective triplets scale with instructions to answer in terms of how they had just been feeling during the stress session. Two other groups that were told to answer in terms of how they had been feeling during the last month and in general, respectively, showed no significant correlations with palmar conductance.

PROCEDURE

Construction of the Forced-Choice Scale

The 20 items from the Taylor *MA* scale which, in independent item analyses by Buss (1955) and Hoyt and Magoon (1954), had been shown to discriminate between criterion groups, were used as the anxiety items in the present scale. These were the same 20 items used by Bendig (1956) in the development of a short form of the Taylor *MA* scale. Twenty-eight other items were selected from the MMPI on the basis of a priori judgment as to their not being directly related to anxiety and their involving personality characteristics that were subject to some fluctuation. The wording of the items was changed, where necessary, so that all items were stated in the past perfect tense. This was done to make it appropriate to answer the items in terms of a specific past time interval.

All 48 items were then rated for social desirability by 110 students from an introductory psychology class on seven-point rating scales. Forty triplets of items were then composed following the format of Heineman (1953) in which an anxiety item was paired with a nonanxiety item of equal social desirability, and a third nonanxiety item was added which differed by approximately two scale units in social desirability (either plus or minus) from the first two items. Each anxiety item appeared twice in the 40 triplets.

The Scoring Procedures

In taking this inventory, subjects were asked to select the item in each triplet that was most like them and the item that was least like them. Scoring

Procedure A was rather complicated and represented an attempt to remove social desirability variance. In brief, the scheme was as follows:

Anxiety item most, nonmatched item least:	3 points
Anxiety item most, nonanxiety nonmatched item least:	2 points
Nonanxiety nonmatched item most, anxiety item not marked:	2 points
Nonanxiety nonmatched item most, anxiety item least:	1 point
Nonanxiety matched item most, anxiety item not marked:	1 point
Nonanxiety matched item most, anxiety item least:	0 points

The logic behind this approach is perhaps best illustrated by examining the 3-point and 0-point combinations. In the former it can be seen that the subject is saying that the matched nonanxiety item is least like him. If putting himself in a favorable or unfavorable light had been important, it is more likely that he would have placed the nonanxiety nonmatched item in either the most or least categories, rather than leaving it in the middle. In the 0-point combination we have the situation in which answering the anxiety item as least like overrides consideration of social desirability since the matched nonanxiety item is marked most like.

Scoring Procedure B consisted simply of giving 2 points if the anxiety item was marked most like, 1 point if left unmarked, and 0 points if marked least like. This score should be influenced considerably by social desirability variance.

The third scoring procedure represented an attempt to measure the social desirability variance itself, although as a result of the nature of the triplet construction, there must inevitably be some negative correlation with the anxiety dimension. The variable was scored as follows, with a high score representing the tendency to say unfavorable things about oneself: 2 points if nonanxiety, nonmatched item was marked least like; 1 point if left unmarked; and 0 points if marked most like.

Subjects and Group Testing

Small groups of volunteer subjects from an introductory psychology course were seen until a total of 40 male and 40 female subjects in each of the three instructional conditions had been administered the Forced-Choice Anxiety scale. The three instructional conditions were obtained by asking the subjects to answer the scale in terms of how they had been during (a) the last 2 weeks, (b) the last 6 months, or (c) in general.

The Individual Stress Session

A random sample of 40 subjects (20 male, 20 female) was selected from each of the larger group tested samples, and contacted for the individual session which occurred on the average about a month after the group session. A more complete description

TABLE 1
CORRELATIONS AMONG THE THREE
CRITERIA MEASURES

Measure	Initial conductance		Rating	
	Male	Female	Male	Female
Systolic change	.31	.13	.28	.30
Initial conductance			.16	.11

Note.— r of .25 is significant at the .05 level for $N = 60$.

of the stress procedure may be obtained in Bergs (1960). Briefly, the subject was confronted in close proximity by two experimenters in a small room. The experimenters watched the subject closely throughout the session and rather obviously made notes and ratings. The subject was told at the beginning,

In this experiment we are going to ask you to do several things. First, we will ask you to tell us what you see on a Rorschach test ink blot. For the second part we will ask you to tell us whatever comes to your mind. We believe that your telling us everything that comes to your mind and your responses to this Rorschach card, together with this apparatus [points to GSR apparatus], will help us understand what your hidden feelings and emotions are, and tell us something about the kind of person you are. But first we want you to sit silently for another couple of minutes before we get started.

Following this anticipation period, the experimenter turned on a tape recorder and presented Rorschach Card II. After the subject had responded, the experimenter commented, "Those responses are not as well integrated as they might be."

The subject was then asked to freely associate for a couple of minutes, during which he was again mildly criticized for not "really" saying everything that came to mind.

Continuous recording of palmar skin conductance and measures of blood pressure at predetermined intervals were obtained during the session. At the end of each session the two experimenters rated the subject on a seven-point scale in terms of how manifestly anxious the subject appeared to be during the session. The correlations between the independent ratings of the two experimenters were .62 for the female subjects and .73 for the male subjects. On the basis of the intercorrelations among the stress measures, a criterion index of anxiety was composed based on the sum of the standard scores for (a) base level conductance, (b) change in systolic blood pressure, and (c) the average of the two experimenters' anxiety ratings. Intercorrelations among the three measures composing the index are shown in Table 1. There were no significant differences between the means of these scores for males and females.

RESULTS AND DISCUSSION

Correlations between the criterion index of anxiety and the Anxiety scale scores for the various instruction groups and scoring procedures are shown in Table 2. The correlations are presented separately for males and females, and it is apparent that the male subjects yielded no significant correlations under any condition. The negative correlational results for the male subjects suggests caution in interpreting the other findings; however, as will be seen the female subjects yield results highly consistent with the theoretical expectations.

For the female subjects the highest correlation, .62, is for the "6-month" instruction group for Scoring Procedure A, the one designed to reduce social desirability variance. The correlation for Scoring Procedure A under "in general" instructions, .49, is also significantly different from zero ($p < .05$) but

TABLE 2
CORRELATIONS BETWEEN THE ANXIETY INDEX AND THE THREE SCORING PROCEDURES
IN THE DIFFERENT GROUPS

Scoring procedure	Instructional time interval					
	In general		6 months		2 weeks	
	Male	Female	Male	Female	Male	Female
A	-.15	.49	-.20	.62	.18	.10
B	.04	.00	-.14	.43	.21	.01
Social desirability	.21	.42	.16	.10	.24	.38

Note.— r of .44 is significant at the .05 level for $N = 20$.

TABLE 3

MEANS AND SIGMAS OF THE THREE SCORING PROCEDURES FOR MALES AND FEMALES COMBINED

Instruction group	Scoring procedure					
	A		B		Social desirability	
	Mean	Sigma	Mean	Sigma	Mean	Sigma
In general	69.88	10.06	37.96	6.16	73.23 ^a	6.50
6 months	69.85	14.21	39.40	11.41	75.33	13.00
2 weeks	72.52	9.93	41.78	8.37	76.32	7.17

^a A high score represents a tendency to admit to socially undesirable characteristics.

not significantly different from the correlation obtained for the "6-month" instruction group. The correlation of .10 for the "2-week" instruction group is significantly less ($p < .05$) than the correlation obtained for the "6-month" group. Thus, as expected, for the female subjects, the "2-week" instruction does have the lowest predictive validity, perhaps because it reflects unduly the more transient states of anxiety. And although the difference between the "6-months" and "in general" correlations is in the expected direction, the difference failed to reach significance.

Scoring Procedure B, where no attempt was made to reduce social desirability variance, did not yield any significant correlations. By employing the somewhat dubious procedure for testing for the significance of the difference between correlations based on the same subjects (McNemar, 1949, p. 124), it was found that the correlations for Scoring Procedure A were significantly higher ($p < .05$) than the correlations obtained with Scoring Procedure B for both the "in general" and "6-months" groups.

None of the correlations for the social desirability scoring procedure was significant, although the correlations were of substantial magnitude for both the "in general" and "2-weeks" groups.

The means and sigmas for the different instruction groups and different scoring procedures are shown in Table 3. There were no significant differences between male and female subjects for any of these means and, accordingly, the two sex groups were combined to yield an N of 80 for each instruction group. It can be seen that there is a general

tendency for the means to increase as the instructional time interval decreases. For Scoring Procedure A this tendency is not significant as tested by analysis of variance. For both Scoring Procedure B and the social desirability scoring procedure, there is a significant effect ($p < .05$) of instructional time interval upon these mean scores. These results are consistent with the theoretical expectation that subjects would admit to more unfavorable characteristics as the time interval decreases. However, one cannot conclude that Scoring Procedure B manifests this effect more than Scoring Procedure A. A correlated-measures analysis of variance was performed on the A and B scoring procedures, and the interaction of scoring procedure by instructional time interval was not found to be significant.

In conclusion the results of the present research indicate that both the instructional time interval and social desirability variance affect the predictive validity of a paper and pencil test of anxiety. It was also found that subjects are likely to say more unfavorable things about themselves when the time interval being reported on is short.

It was not the purpose of this paper to publish a new psychometric test and, in fact, item analyses of the present scale (not reported in this paper) suggest that many of the item triplets are not predictive at all. The completely negative results for the male subjects emphasize this point.

SUMMARY

The primary purpose of this research was to study the effect of instructional time in-

terval and social desirability variance upon the validity of a forced-choice anxiety scale. Subjects in three different groups were asked to answer the scale in terms of how they had been during the last 2 weeks, last 6 months, and in general. The forced-choice triads were then scored by Procedure A, which was designed to reduce variance associated with the social desirability of the items, and by Procedure B, which was presumed to be heavily affected by the social desirability of the items. A criterion index of anxiety was obtained in an individual stress session, and was based on skin conductance level, change in systolic blood pressure, and a rating of anxiety.

The results were entirely negative for the male subjects; no significant correlations were found for any instruction group or for either scoring procedure. For the female subjects the results were in accord with the theoretical expectations. The highest predictive validity was obtained for the "6-month" instruction group for the scoring procedure that was designed to minimize social desirability variance. The correlation with the criterion was also significant for the "in general" instruction group for Scoring Procedure A. No criterion correlations were significant for Scoring Procedure B.

A second aim of the research was to study the effects of instructional time interval upon

the mean social desirability scores, which were assessed by a third scoring procedure. As expected, subjects tended to admit to more socially undesirable characteristics as the instructional time interval decreased.

REFERENCES

- BENDIG, A. W. The development of a short form of the manifest anxiety scale. *J. consult. Psychol.*, 1956, 20, 384.
- BERGS, L. P. The validity of a revised form of the forced-choice manifest anxiety scale as a function of the time interval being assessed. Unpublished master's thesis, University of Wisconsin, 1960.
- BUSS, A. H. A follow up item analysis of the Taylor anxiety scale. *J. clin. Psychol.*, 1955, 11, 409-410.
- EDWARDS, A. L. *The social desirability variable in personality assessment and research*. New York: Dryden, 1957.
- HEINEMAN, C. E. A forced-choice form of the Taylor anxiety scale. *J. consult. Psychol.*, 1953, 17, 447-454.
- HOYT, D. P., & MAGOON, T. M. A validation study of the Taylor manifest anxiety scale. *J. clin. Psychol.*, 1954, 10, 357-361.
- McNEMAR, Q. *Psychological statistics*. New York: Wiley, 1949.
- MARTIN, B. The validity of a self-report measure of anxiety as a function of the time interval covered by the instructions. *J. consult. Psychol.*, 1959, 23, 468.
- SILVERMAN, R. E. The manifest anxiety scale as a measure of drive. *J. abnorm. soc. Psychol.*, 1957, 55, 94-97.

(Received October 7, 1960)

RESPONSE STYLES IN CLINICAL AND NONCLINICAL GROUPS

H. J. WAHLER

University Health Center, Ohio State University

Over the past decades, conceptions of personality measurement have undergone various transitions. One change is reflected in an increased interest in molar response characteristics conceived as response sets (Cronbach, 1946), or response styles (Jackson & Messick, 1958), elicited by the verbal stimuli of questionnaires. Two response sets have been of particular interest currently. One, Edwards (1953) describes as the "social desirability variable." He and other investigators found high correlations in the vicinity of .87 between rated item social desirability and averaged scores of student groups on self-descriptive questionnaire items. The other is a response set which Cronbach (1946, 1950) termed acquiescence or a tendency to agree (or disagree) with items irrespective of their content. In their recent review, Jackson and Messick (1958) conclude that:

In the light of accumulating evidence, it seems likely that the major common factors in personality inventories of the true-false or agree-disagree type . . . are interpretable primarily in terms of style rather than specific item content (p. 247).

If the major common factors in personality inventories are interpretable in terms of response (R) styles, then two groups which differ significantly on a scale of psychiatric symptomatology should also differ significantly in terms of R styles shown by the members. A sample of patients that cannot be so discriminated from controls should not show different proportions of R styles than controls. Also subgroups of subjects from different clinical and nonclinical groups who exhibit the same R styles on one questionnaire should have comparable scores on different scales. Furthermore, if R styles may be interpreted as the major common factor rather than specific item content, scores obtained by subjects with one set of personality items

should covary with their scores derived from different scales with different content, the direction being consistent with the bias of their R style.

Messick and Jackson also point out that R set studies have tended to focus on one or another R style such as acquiescence or social desirability without studying both conjointly. One point which particularly bears special attention is the possibility that the set to agree or disagree may interact with item desirability.

The purpose of this study is to investigate the above propositions which may be briefly restated as four questions: (a) Are significant differences found between clinical and non-clinical groups in terms of the frequency with which different R styles occur when the clinical group can be differentiated from controls by a scale of general psychiatric symptomatology, and when clinical and control groups cannot be so discriminated? (b) Do subjects who show the same R styles in self-ratings obtain comparable scores with true-false scales in spite of their being members of different clinical and nonclinical groups? (c) Do subjects exhibit the same R sets with different items and modes of responding? For example, if they tend to deny traits on a self-rating scale do they show the same tendency with another set of items and a true-false mode of responding? (d) Is the tendency to claim or deny undesirable traits related to claiming or denying desirable traits or are these tendencies independent? Do the clinical and control samples differ in this respect?

PROCEDURE

Response Styles

Couch and Keniston (1960) have shown a significant correspondence between average level of response to items rated on a seven-point scale and the number of "true" responses to MMPI items. This

they conclude shows that the R set to agree or disagree is demonstrable with both methods of responding. If this is the case, four different R styles may be readily defined based on the level of individual subjects' responses to a self-rating inventory. These indices reflect the three response sets of major current interest, namely, the set to agree, to disagree, and to give responses that correlate positively with perceived social values associated with items, and a fourth style which is the opposite of the latter. With a self-rating inventory containing items judged desirable (D) and other items judged undesirable (U), a two-way classification of scores in terms of level (high-low mean ratings) and item desirability (D-U) can serve to define the four R styles. A low score on either D or U scales was defined as one below the median of the distribution of scores for all subjects combined. A high score was defined as lying above the median of combined distributions. Subjects who rate low on both D and U scales are showing a tendency to deny or disagree irrespective of content and perceived social values of items. Subjects rating high on both D and U scales are exhibiting an R style indicative of an agreeing set. Subjects who rate high on D and low on U are, by definition, manifesting a social desirability R style. The fourth style evolves logically from the two-way classification of scores. Subjects rating low on D and high on U scales would fall in this class. Logically this R style can be described as the antithesis of a social desirability set, a social *undesirability* set.

Self-Description Inventory

The self-description inventory (SDI) contains 44 items with heterogeneous content pertaining to characteristics of common clinical interest such as anxiety, hostility, sexual adjustment, interpersonal adjustment, dependency, etc. This content is phrased in the first person with nontechnical language, i.e., "I have trouble getting along with people." Subjects are instructed to rate themselves on these items by means of a nine-point scale with anchoring statements ranging from "not at all like me" to "beyond question very much like me."

Twenty-seven of the items which had been independently rated as slightly to highly undesirable (mean ratings of less than five with a nine-point scale of D-U) were selected in conjunction with a different study, Wahler (1958), on the basis of their ability to discriminate mental hygiene intake and nonpatient groups at the .05 or less level of significance. These 27 items comprise the U traits scale used in classifying R styles. Eight items which were judged slightly to highly desirable (mean ratings of greater than 5) make up the D traits scale. The U score for each subject is the mean of his self-ratings on the U items and the D score is the mean of his self-ratings on the D items.

MMPI Scales

Three MMPI scales were selected as measures in this study since they contain a variety of content and require a true-false mode of responding which

differs from the self-rating approach with the SDI. Norma Besch and James Taylor kindly made data collected by them available to the author. Besch had 71 male undergraduates at Ohio State University rate 200 MMPI items for "personal desirability" on a nine-point scale; the *Pt* and *L* scales were included among the items. Taylor obtained social desirability ratings from 81 adult normals on 205 MMPI items which included the *K* scale. From these ratings it was evident that items making up the *K*, *L*, and *Pt* scales are perceived as primarily undesirable. Seventy-three percent of the *K* items were judged undesirable. The 22 undesirable *K* items had a mean median rating of 3.79 on a nine-point scale. Eighty-one percent of the 48 *Pt* items were judged undesirable with a mean mean rating of 3.28 for the 39 undesirable items. Eighty percent of the *L* items were judged U with a mean mean rating of 4.13. Scores from each of these scales accrue mainly from responses in one direction. The *K* score increases with the number of items denied except for 8 out of 30. The magnitude of the *L* score is also a function of denying items in all cases but 3 out of 15. *Pt* scores, on the other hand, are mainly a function of the number of items claimed or agreed with; this is true for 39 out of 48 items. Furthermore, the item overlap is negligible among these three scales with only one common item (J-51) scored in the same direction on the *K* and *L* scales.

Subjects

The nonpsychiatric subjects consisted of 26 male and 44 female sophomores taking an introductory psychology course and 39 male students taking an elementary personality course at the State University of Iowa. The SDI and a "Biographical Inventory" containing 240 MMPI items were administered to these people in groups. These people were told that their responses were to be studied as part of a research project and would not be used in any personal way.

Two different subpopulations of clinical subjects were sampled on the assumption that outpatients who voluntarily seek help at a mental hygiene clinic are more likely to describe their characteristics frankly (as they conceive them) than are hospitalized patients who as a group often tend to exhibit more severe pathology manifested by extreme reality distortion and resistive defenses exhibited as marked denial or negativism who frequently have been pressured by relatives and/or community to consign themselves to conditions which they may not like and/or believe they don't need and hope to leave by appearing "normal."

The Mental Hygiene Clinic (MHC) subjects (assumed more frank) consisted of 47 males in the process of applying for help with personal problems at a Veterans Administration Mental Hygiene Clinic. Their mean age was 32.6 with a range of 22 to 48. All tests were administered individually in the course of their evaluation for treatment. Subjects in this group were given the card form of the MMPI. Also, a briefer form of the SDI was given this group; the U score is based on 22 of the 27 items and the D score is based on 5 of the 8 desirable items.

TABLE 1
MEAN SCORES OF GROUPS ON MMPI SCALES AND SELF-RATINGS

Subjects	N	SDI		MMPI					
		D	U	K	Pt	L			
Student groups combined (A)	109	5.15	3.11	15.99	13.06	2.85			
MHC (B)	47	5.61	4.94	12.55	21.36	3.47			
VAH ^a (C)	75	5.47	3.23	19.67	13.92	6.08			
Pairs of scores compared**		<i>t</i>	<i>p</i>	<i>t</i>	<i>p</i>	<i>t</i>	<i>p</i>	<i>t</i>	<i>p</i>
A:B		2.5	.025	8.4	.001	4.5	.001	5.7	.001
A:C		2.1	.05	<i>ns</i>		3.8	.001	<i>ns</i>	6.6 .001
B:C		<i>ns</i>		7.4	.001	6.5	.001	3.6	.001

^a All MMPI Scores for VAH group based on *N* of 24.

** *F* tests over all column means were all significant at the .025 or less level.

The hospitalized (VAH) group (assumed more guarded) consisted of 75 patients, 65% of whom were diagnosed as some type of schizophrenic; the remaining were diagnosed as other types of functional psychiatric disorders. Their mean age was 34.4 with all but two cases within the range of 18 to 49. The two exceptions were 60 and 61 years of age. All subjects were administered the SDI individually during admission testing. Tests other than the SDI were administered at the discretion of the examiners; only 24 were given the card form of the MMPI.

RESULTS

The assumption that patients seeking help at an outpatient clinic respond more frankly (less defensively) than hospitalized patients

as a group was well substantiated by the relative levels of mean *Pt*, *K*, and *L* scores of these groups in comparison with each other and controls as shown in Table 1. MHC subjects had significantly higher *Pt* scores than controls or hospitalized patients while the mean *Pt* of VAH subjects did not differ significantly from controls. To the extent that the *Pt* reflects psychiatric symptomatology, the hospitalized patients claimed no more symptoms than did students. Thus, the VAH group qualifies as a clinical sample that is not differentiated from controls by a scale reflecting psychiatric pathology while the MHC

TABLE 2
CONTINGENCY ANALYSIS OF FREQUENCY OF FOUR R STYLES SHOWN BY GROUPS

Group	R styles denying		Socially undesirable		Socially desirable		Agreeing	
	Low D-Low U		Low D-High U		High D-Low U		High D-High U	
VAH	21	(28)	9	(12)	24	(32)	21	(28)
MHC	3	(6)	19	(40)	1	(2)	24	(51)
Student (F)	18	(42)	8	(18)	15	(34)	3	(7)
Student (M)	18	(28)	15	(23)	16	(25)	16	(25)

Over-all $\chi^2 = 51.4$, $p < .001$

Group comparisons

VAH:MHC, $\chi^2 = 33.8$, $p < .001$

VAH:F, $\chi^2 = 8.4$, $.02 < p < .05$

VAH:M, $\chi^2 = 3.3$, $p > .30$

MHU:F, $\chi^2 = 43.7$, $p < .001$

MHU:M, $\chi^2 = 23.7$, $p < .001$

F:M, $\chi^2 = 7.3$, $.05 < p < .10$

Note.—Values in parentheses are percentages.

TABLE 3
MEAN *Pt* SCORES OF GROUPS CLASSED BY R STYLES

Group	R styles denying		Socially undesirable		Socially desirable		Agreeing		<i>F</i>	<i>df</i>	<i>p</i>
	<i>M</i>	<i>N</i>	<i>M</i>	<i>N</i>	<i>M</i>	<i>N</i>	<i>M</i>	<i>N</i>			
VAH	12.2	8		1	11.7	9	20.7	6	1.3	2/20	<i>ns</i>
MHC	6.0	3	20.8	19		1	24.0	24	5.4	2/43	.01
Student (F)	11.1	18	15.5	8	8.7	15	18.7	3	4.8	3/40	.01
Student (M)	9.7	18	16.0	15	11.0	16	20.1	16	11.1	3/61	.001
<i>F</i>	0.8		2.0		0.8		0.8				
<i>df</i>	3/43		2/39		2/37		3/45				
<i>p</i>	<i>ns</i>		<i>ns</i>		<i>ns</i>		<i>ns</i>				

group was so differentiated. Also, the VAH patients scored significantly higher than MHC or student subjects on *K* and *L* indices of defensiveness.

The most striking difference in frequencies of R styles is exhibited by the MHC group; as shown in Table 2, 91% of the MHC subjects had socially undesirable or agreeing styles. The over-all differences in frequencies within Table 1 are highly significant ($p < .001$). Comparisons of groups in pairs show that the MHC group differs significantly from all others. Frequencies of R styles shown by the VAH group did not differ significantly from those of control males but did differ significantly from those of females ($.02 < p < .05$).

When these frequencies are grouped in terms of high or low scores on U, disregarding D, the group differences remain significant. When the frequencies are classed in

terms of high or low scores on D, disregarding U, differences between groups are no longer significant.

Subjects from the different diagnostic groups were formed into subgroups based on the R styles they showed in responding to the SDI. Mean *Pt*, *K*, and *L* scores were computed for each of these subgroups. In Table 3 it may be seen that the mean *Pt* scores within each column are quite comparable for members of different diagnostic groups with the same R styles. None of the *F* tests for significance of differences within columns attained significance at the .05 level.

When row means are compared (within diagnostic group across R styles) it is evident that within each of the four diagnostic groups, subjects who had socially undesirable or agreeing R styles obtained higher mean *Pt* scores than did those with denying or social desirability styles. Differences between these

TABLE 4
MEAN *K* SCORES OF GROUPS CLASSED BY R STYLES

Group	R styles denying		Socially undesirable		Socially desirable		Agreeing		<i>F</i>	<i>df</i>	<i>p</i>
	<i>M</i>	<i>N</i>	<i>M</i>	<i>N</i>	<i>M</i>	<i>N</i>	<i>M</i>	<i>N</i>			
VAH	19.8	8		1	21.4	9	16.7	6	1.54	2/20	<i>ns</i>
MHC	17.3	3	13.0	19		1	11.8	24	2.8	2/43	<i>ns</i>
Student (F)	18.3	18	14.6	8	17.6	15	11.7	3	4.1	3/40	.05
Student (M)	18.6	18	13.5	15	16.9	16	12.8	16	7.4	3/61	.001
<i>F</i>	.4		.4		7.0		2.4				
<i>df</i>	3/43		2/39		2/37		3/45				
<i>p</i>	<i>ns</i>		<i>ns</i>		.01		<i>ns</i>				

TABLE 5
MEAN *L* SCORES OF GROUPS CLASSED BY R STYLES

Groups	R styles denying		Socially undesirable		Socially desirable		agreeing		<i>F</i>	<i>df</i>	<i>p</i>
	<i>M</i>	<i>N</i>	<i>M</i>	<i>N</i>	<i>M</i>	<i>N</i>	<i>M</i>	<i>N</i>			
VAH	5.6	8		1	6.4	9	5.8	6	0.1	2/20	<i>ns</i>
MHC	6.3	3	3.3	19		1	3.2	24	3.6	2/43	.05
Student (F)	3.2	18	3.2	8	3.5	15	3.0	3	0.1	3/40	<i>ns</i>
Student (M)	2.8	18	2.7	15	2.4	16	2.4	16	0.2	3/61	<i>ns</i>
<i>F</i>	6.0		0.5		8.7		2.8				
<i>df</i>	3/43		2/39		2/37		3/45				
<i>p</i>	.01		<i>ns</i>		.001		<i>ns</i>				

means for groups with significant *F* tests across R styles were significant at the .05 or less level, as determined by *t* tests of pairs of means, with one exception where the *Pt* mean in the socially undesirable category was not significantly different from that in the denying category for female students. None of the mean *Pt* scores of subjects exhibiting denying or social desirability R styles differed significantly; this was also true for subjects showing socially undesirable or agreeing styles.

The same analyses were computed for *K* scores. In this case, there were no significant differences within R styles (columns) except in the social desirability category. Here, the VAH group obtained significantly higher *K* scores than did either of the student groups. In Table 4 it may be seen that subjects who had socially undesirable or agreeing R styles irrespective of diagnostic group membership obtained lower mean *K* scores than did subjects in the social desirability or denying categories. Again, *t* tests of differences between combinations of *K* means taken two at a time

for the two student groups (groups with significant *F*s across R style categories) yielded significant differences at the .05 or less level between *K* means obtained from subjects with denying or social desirability styles and those with socially undesirable or agreeing sets.

Comparable analyses of the mean *L* scores showed different trends than were found with *K* or *Pt*. It may be seen in Table 5 that, in general, VAH patients have the highest *L* scores irrespective of R styles except for the small number of MHC cases who exhibited the denying style. There were no significant differences between the two student groups either across sex classifications or across R style categories. The mean *L* scores of MHC cases in the socially undesirable and agreeing categories do not differ significantly from those of students. The VAH group had a significantly higher mean *L* than either student group in the social desirability and denying categories; the MHC *L* was also significantly higher than mean *L* scores of students in the latter category.

TABLE 6
CORRELATIONS AMONG MMPI AND SELF-RATING SCORES OF GROUPS

Group	<i>N</i>	Scale									
		<i>K:L</i>	<i>K:Pt</i>	<i>K:U</i>	<i>K:D</i>	<i>Pt:L</i>	<i>Pt:U</i>	<i>Pt:D</i>	<i>L:U</i>	<i>L:D</i>	<i>U:D</i>
VAH	24	.59*	-.50*	-.41*	-.15	-.06	.35	.26	.15	.44*	.34 ^a *
MHC	47	.31*	-.57*	-.36*	-.28	-.45*	.80*	.22	-.43*	-.13	.04
Student (F)	44	.31*	-.63*	-.55*	.01	-.37*	.58*	-.34*	-.22	.19	-.31*
Student (M)	65	.27*	-.75*	-.69*	-.06	-.24*	.72*	.24*	-.16	-.18	.14

^a *N* = 75 in this instance for VAH group.

* Significant at .05 or less level of significance.

The intercorrelations of *K*, *L*, *Pt* scores and mean *U* and *D* ratings are presented in Table 6. Nineteen out of the 24 correlations among scales composed mainly or entirely of undesirable items (*K*, *L*, *Pt*, and *U*) are statistically significant at the .05 or less level and in all instances except one (*L:U*) the coefficients are in the same direction across groups. The magnitudes and directions of these coefficients indicate that subjects tend to deny or claim socially undesirable traits with some degree of consistency on different scales. The mean ratings on desirable items were correlated with scores from the scales comprised of undesirable items; only 5 out of the 16 coefficients were significant and the directions of the coefficients across groups were not consistent.

DISCUSSION

The assumption that a "frank" clinical group should exhibit significantly different proportions of *R* styles relative to controls was supported by the findings. It was also found that the various *R* styles occurred with about the same frequency in a comparable control group and a group of clinical cases that could not be discriminated by a scale of psychiatric symptomatology. Both of these findings are in accord with Jackson and Messick's conclusion that the major common factors in personality inventories are interpretable in terms of response style.

From their conclusion it would also be predicted that subgroups of subjects from different clinical and nonclinical groups who exhibit the same *R* styles on one questionnaire should have comparable scores on different scales. For two of the MMPI scales this was found to be the case; with the *L* scale, hospitalized subjects scored consistently higher than other subjects, irrespective of *R* styles. Eichman (1959), in comparing female schizophrenic patients with controls, also found that high *L* scores characterized his samples of hospitalized patients.

It was found that some scales with different content reflect the bias of the respondent's *R* style in a consistent direction, but not others.

The correlational analyses provide evidence that members of all groups tended to respond with some degree of consistency to scales com-

posed of socially undesirable items. Correlations of the various scales with mean self-ratings on *D* items failed to show any consistent trends and the majority of the coefficients were not significant. These findings are in opposition to an assumption that a general acquiescence set may operate independently of the judged desirability or undesirability of items. In general, the above findings constitute or suggest interactions.

It was not possible in this study to use a four factor analysis of variance design because the limited number of clinical subjects made it infeasible to discard cases in order to achieve the proportionality of scores required for such analyses. However, four factors are represented in the more fragmented series of analyses: (a) samples from different subpopulations (this group's factor is confounded with situational factors); (b) different scales, and two categorical variables used in classifying subjects; (c) high or low scores (claiming or denying tendencies) on self-rating scales; and (d) classifications of items as socially desirable or undesirable on the basis of independent ratings.

The essential *R* style factor found in this study was what DeSoto and Kuethe (1959), term the "symptom-claiming set"—a disposition on the part of subjects to claim (or deny) undesirable symptomatic traits. But this factor is hardly pure since it interacts with scales—which may in turn be a function of different types of content. A Type I design (Lindquist, 1953) was used to test the significance of the scales by *R* styles interaction; this interaction was significant ($p < .001$) comparing MMPI scores on *K*, *L*, and *Pt* for all subjects classed as denying undesirable traits on self-ratings and subjects tending to claim undesirable traits. It was also possible with this material to analyze the group by scales interaction, which likewise was significant at less than the .001 level. Analyses of frequency of *R* styles across groups showed an item desirability by groups interaction. While a more elegant complete factorial analysis was not possible, the evidence clearly points toward multiple interaction effects among the variables considered.

If the general implications of these findings are borne out by additional and more explicit

evidence, Jackson and Messick's statement might need rephrasing. For example, it might have to be changed to read "it seems likely that the major common factors in personality inventories of the true-false or agree-disagree type . . . are interpretable primarily in terms of . . ." R styles which in turn are functions of interacting variables such as social desirability of items, certain specific types of item content, differential characteristics of the subpopulations sampled, and the circumstances under which subjects are tested.

SUMMARY

Several implications follow from the proposition that the major common factors in personality inventories are interpretable mainly in terms of response styles. Among them are the following three hypotheses: (a) Groups which differ significantly in terms of scores on a scale of symptomatology should also differ in the frequency with which various R styles are shown by the members, and that groups which cannot be differentiated on the basis of such scores should not exhibit significantly different frequencies of R styles. (b) Subjects who exhibit the same R styles on one instrument should have similar scores on different scales even though they are members of different subpopulations. (c) Subjects' R sets should be manifested in a consistent direction on different scales. A question may also be raised as to whether R sets operate independently of or interactively with the social desirability values associated with items comprising scales.

The findings show that a group of "frank" outpatient subjects exhibit a significantly different frequency of R styles than controls and "guarded" patients. A group of "guarded" hospitalized patients who scored at the same level as controls on a scale of symptomatology could not be discriminated from controls by differential frequency of R styles. Subjects from different diagnostic groups classed according to the R styles they showed on self-ratings had similar mean *Pt* and *K* scores

within each R style category but not similar *L* scores. Subjects' tendencies to deny or claim undesirable characteristics were exhibited relatively consistently in terms of mean scores on self-ratings and the *Pt* and *K* scales. Different diagnostic groups responded differently to the *L* scale irrespective of R styles. Scores on scales comprised of items describing undesirable traits were found to covary in a consistent direction. No such consistency was found when these scores were correlated with a scale composed of desirable items.

Consistent response characteristics interpretable as R sets or styles were found, but it was inferred from the findings that multiple interaction effects are likely among variables such as (a) the particular type of R style shown, (b) subpopulation differences, (c) desirability and undesirability of items comprising scales, and (d) other scale differences such as types of content.

REFERENCES

- COUCH, A., & KENISTON, K. Yeasayers and naysayers: Agreeing response set as a personality variable. *J. abnorm. soc. Psychol.*, 1960, 60, 151-174.
- CRONBACH, L. J. Response sets and test validity. *Educ. psychol. Measmt.*, 1946, 6, 616-623.
- CRONBACH, L. J. Further evidence on response sets and test designs. *Educ. psychol. Measmt.*, 1950, 10, 3-31.
- DESOTO, C. B., & KUETHE, J. L. The set to claim undesirable symptoms in personality inventories. *J. consult. Psychol.*, 1959, 23, 496-500.
- EDWARDS, A. L. The relationship between the judged desirability of a trait and the probability that the trait will be endorsed. *J. appl. Psychol.*, 1953, 37, 90-93.
- EICHMAN, W. J. Discrimination of female schizophrenics with configural analysis of the MMPI profile. *J. consult. Psychol.*, 1959, 23, 442-447.
- JACKSON, D. M., & MESSICK, S. Content and style in personality assessment. *Psychol. Bull.*, 1958, 55, 243-252.
- LINDQUIST, E. F. *Design and analysis of experiments in psychology and education*. Boston: Houghton Mifflin, 1953.
- WAHLER, H. J. Social desirability and self-ratings of intakes, patients in treatment, and controls. *J. consult. Psychol.*, 1958, 22, 357-363.

(Received October 12, 1960)

THE CLINICAL UTILITY OF "INVALID" MMPI *F* SCORES

MALCOLM D. GYNTHIER

Washington University Medical School

Many investigators (e.g., Astin, 1959; Gilbertstadt & Duker, 1960; Rempel, 1958) eliminate MMPI profiles containing high *F* scores from analyses on the grounds that such profiles are not valid. This procedure is consistent with early injunctions to omit such profiles from research work (Hathaway & Meehl, 1951), but inconsistent with more recent views or experimental findings which emphasize the characterological implications of scores on the validity scales (Gough, 1956b; Gross, 1959), rather than test taking attitudes as such. Determination of the relationship between high *F* scores, diagnostic classification, and aggressive versus passive criminal behavior would seem to be helpful in demonstrating whether such "invalid" *F* scores have any predictive value.

METHOD

Test data of all 353 white male court referrals admitted to South Carolina State Hospital between September 1957 and August 1960 were examined. Two hundred fifty-one completed MMPIs (all cases testable by this method) were found. Five profiles given by organic patients were excluded because this number of cases was insufficient for analysis as a separate category. The remaining 246 cases were sorted into subsamples according to the diagnostic impression of the psychiatric staff, which was not based on the MMPI data. This procedure yielded 194 behavior disorders (BD), 29 neurotics (N), and 23 psychotics (P). Intelligence estimates in the form of Kent-EGY, Scale D scores were available for all cases. Means for the BD, N, and P groups were 28.16, 27.14, and 27.43, respectively. (The average range is 24-31, inclusive.) Statistical analyses revealed no significant differences between the groups which suggests that whatever differences there are between distributions of *F* scores cannot be attributed to differences in intelligence. Mean ages for the BD, N, and P groups were 30.31, 39.96, and 37.83 years, respectively. Statistical analyses showed that the BD group is significantly younger ($p < .01$) than either of the other groups.

Different investigators use different *F* values as a basis for discarding data. Sometimes the reader is only informed that cases were removed because the validity scores were "high" (e.g., Rosen, 1958; Sopchak, 1958), but in most cases the exact cutting score is given (e.g., Goodstein & Dahlstrom, 1956; Pantton, 1958). In this investigation, high was defined as $F > 16$ raw score points, following the recommendation of Gough (1956a) and Meehl (1956).

RESULTS AND DISCUSSION

Table 1 shows the distribution of *F* scores for the BD, N, and P groups. Mean *F* scores were 8.66, 6.76, and 8.04, respectively. Statistical analyses revealed no significant differences which implies that differences in the total distribution of *F* scores cannot be attributed to differences in diagnostic classification. However, there were striking differences between the groups with regard to distribution of *F* scores greater than 16. Thirty-nine scores fell into this invalid category, with 37 of these being given by individuals classified as behavior disorders. Percentages of $F > 16$ scores for the BD, N, and P groups were 19.1, 0, and 8.7, respectively. Chi square, corrected for continuity, showed that these differences depart significantly from chance ($\chi^2 = 6.04$, $df = 2$, $p < .05$).

The significant age differences between the groups raise the question of whether the differences in $F > 16$ distributions might not be explained by the age differences alone. Analysis of the younger and older halves of the BD group showed that the younger subsample gave $F > 16$ score, more frequently than the older men ($\chi^2 = 5.64$; $df = 1$, $p < .02$). Also, the two $F > 16$ scores found in the psychotic group were given by 22-year-old men. Age is obviously an important factor. However, if the mean age of the BD group is adjusted (by eliminating every other subject 30 years

old or younger) so that it is not significantly different from the P group, the BD group still had twice the percentage of $F > 16$ scores than the P group (25/141 or 17.7% versus 8.7%).

It would appear that invalid MMPI F scores can discriminate between diagnostic classifications. That is, in groups of male court referrals matched for age and intelligence, behavior disorders obtained 67% of the $F > 16$ scores, psychotics 33%, and neurotics 0%. And, if one were to consider only individuals 23 years of age or older, 100% of the $F > 16$ scores would be obtained by behavior disorders.

These court-referred individuals differ from psychiatric patients-in-general in that they all have a reason for "faking bad": to decrease the probability that they will be convicted of the crimes with which they are charged. It would be worthwhile to replicate this study with routinely admitted psychiatric patients to see if our results are substantiated by a group with less reason for dissembling. One check for dissembling in these court-referred subjects is to test the hypothesis that faking bad is positively related to the severity of the crime with which the person is charged. Analysis of the data given by murderers and rapists ($N = 31$) does not support the hypothesis, since the percentage of $F > 16$ scores given by this subgroup with the most serious charges against them is nearly equivalent to the percentage obtained by the remainder of the sample (16.1 versus 15.8). Furthermore, the dissembling interpretation does not account for the differential $F > 16$ results found with the different diagnostic classes.

With regard to the characterological interpretation of the F scale, it is interesting to note that Leary (1956) considers F to be a measure of the aggression and sadism to be expected in interpersonal relations. Thus, the higher the elevation on F , the more cruel and unkind the individual is predicted to behave. From that point of view, our subjects who obtained $F > 16$ scores would be considered as more aggressive in an antisocial manner than the remainder of the sample. Analysis of the $F > 16$ scores obtained by those who committed aggressive crimes (i.e., stealing,

TABLE 1
FREQUENCY DISTRIBUTION OF RAW SCORES ON THE
MMPI F SCALE FOR BEHAVIOR DISORDER, NEUROTIC
AND PSYCHOTIC GROUPS

Raw score	BD	N	P
35-37	0	0	1
32-34	4	0	0
29-31	8	0	0
26-28	2	0	0
23-25	10	0	0
20-22	7	0	1
17-19	6	0	0
14-16	8	2	2
11-13	11	3	1
8-10	19	8	5
5-7	25	4	1
2-4	66	10	10
0,1	28	2	2
<i>N</i>	194	29	23

rape, murder, etc.) versus those who committed passive crimes (i.e., forgery, breach of trust, drunken driving, etc.) shows that there is a tendency for the aggressive criminals to obtain $F > 16$ scores more frequently than the passive criminals ($\chi^2 = 3.04$, $df = 1$, $.05 < p < .10$).

The interpretation of MMPI F scores as indicating aggressiveness also casts some light on the differential $F > 16$ scores by diagnosis. Neurotics, who obtained no $F > 16$ scores, tended to commit passive or asocial crimes, whereas the behavior disorders and psychotics tended to commit antisocial crimes. An illustration may clarify this point. Of the 14 sex crimes committed by neurotics, 8 were incest, 3 were rape, and 3 were "Peeping Tom." With regard to the 42 sex crimes committed by behavior disorders and psychotics, 21 were rape or attempted rape, 7 were lewd acts on children, 6 were indecent exposure, 5 were incest, 2 were sending obscene letters to women, and 1 was "Peeping Tom." This latter group would appear to contain a far higher percentage of aggressive acts against society than the neurotic group, which is consistent with the differential $F > 16$ findings and the interpretation of F as an indicator of aggressive and sadistic behavior.

SUMMARY

This study investigated the relations between "invalid" MMPI F scores, diagnostic classes, and aggressive versus passive criminal behavior to determine if $F > 16$ scores, which usually lead to elimination of the MMPI prior to analysis of the data, have any predictive significance. MMPIs were available from 246 white male court referrals who were classified as behavior disorder ($N = 194$), neurotic ($N = 29$), or psychotic ($N = 23$) by the psychiatric staff.

Thirty-nine of the 246 subjects obtained $F > 16$ scores. Thirty-seven of these 39 deviant scores were obtained by behavior disorders. When the data were adjusted to equate the groups on age and intelligence, behavior disorders were shown to have 67% of the $F > 16$ scores, psychotics 33%, and neurotics 0% of such scores. It was also demonstrated that younger men more frequently obtain invalid F scores than older men. Although all the subjects had reason to dissemble, the results seem most consistent with a characterological interpretation of the F scale.

The practice of discarding MMPI data because of invalid F scores seems highly questionable, especially if the investigator wishes to draw valid conclusions about groups, such as behavior disorders, who are likely to display aggressive, antisocial actions.

REFERENCES

- ASTIN, A. W. A factor study of the MMPI psychopathic deviate scale. *J. consult. Psychol.*, 1959, 23, 550-554.
- GILBERSTADT, H., & DUKER, J. Case history correlates of three MMPI profile types. *J. consult. Psychol.*, 1960, 24, 361-367.
- GOODSTEIN, L. D., & DAHLSTROM, W. G. MMPI differences between parents of stuttering and non-stuttering children. *J. consult. Psychol.*, 1956, 20, 365-370.
- GOUGH, H. G. Diagnostic patterns on the MMPI. In G. S. Welsh & W. G. Dahlstrom (Eds.), *Basic readings on the MMPI in psychology and medicine*. Minneapolis: Univer. Minnesota Press, 1956. Pp. 340-350. (a)
- GOUGH, H. G. The F minus K dissimulation index for the MMPI. In G. S. Welsh & W. G. Dahlstrom (Eds.), *Basic readings on the MMPI in psychology and medicine*. Minneapolis: Univer. Minnesota Press, 1956. Pp. 5-11. (b)
- GROSS, L. R. MMPI L - F - K relationships with criteria of behavioral disturbance and social adjustment in a schizophrenic population. *J. consult. Psychol.*, 1959, 23, 319-323.
- HATHAWAY, S. R., & MEEHL, P. E. *An atlas for the clinical use of the MMPI*. Minneapolis: Univer. Minnesota Press, 1951.
- LEARY, T. F. *Multilevel measurement of interpersonal behavior*. Berkeley, Calif.: Psychological Consultation Service, 1956.
- MEEHL, P. E. Profile analysis of the MMPI in differential diagnosis. In G. S. Welsh & W. G. Dahlstrom (Eds.), *Basic readings on the MMPI in psychology and medicine*. Minneapolis: Univer. Minnesota Press, 1956. Pp. 292-297.
- PANTON, J. G. MMPI profile configurations among crime classification groups. *J. clin. Psychol.*, 1958, 14, 305-308.
- REMPEL, P. P. The use of multivariate statistical analysis of MMPI scores in the classification of delinquent and nondelinquent high school boys. *J. consult. Psychol.*, 1958, 22, 17-23.
- ROSEN, A. Differentiation of diagnostic groups by individual MMPI scales. *J. consult. Psychol.*, 1958, 22, 453-457.
- SOPCHAK, A. L. Spearman correlations between MMPI scores of college students and their parents. *J. consult. Psychol.*, 1958, 22, 207-209.

(Received October 20, 1960)

INTERACTION OF BRAIN INJURY WITH PERIPHERAL VISION AND SET

HAROLD L. WILLIAMS, CHARLES F. GIESEKING, AND ARDIE LUBIN

Walter Reed Army Institute of Research

On tests such as Kohs' Block Design, the Bender-Gestalt, or Benton's Memory for Designs where designs have to be reproduced, the phenomenon called rotation frequently occurs. The subject reproduces the design correctly, but tilts it at an angle to the target design, sometimes as much as 45° to 90° . Rotation has been observed frequently in brain injured patients, in children, and in dull normals (Bender & Teuber, 1948; Goldstein & Scheerer, 1941; Hanvik & Anderson, 1950; Pascal & Suttell, 1951).

The Block Design Rotation Test (BDRT¹), devised by Shapiro (1951), was used in a series of studies by Shapiro (1951, 1952, 1953) and Yates (1954) to show that: geometric properties of the target design had a significant effect on rotation, intelligence correlated negatively with rotation, reducing peripheral vision increased rotation in normal subjects. Williams, Lubin, Giesecking, and Rubinstein (1956) confirmed these findings but found that much of the rotation variance was accounted for by intelligence. This effect was so strong that dull normal subjects could not be discriminated from brain injured on the basis of their rotation scores. In addition, they found an interaction between brain injury and peripheral vision; restricting peripheral vision did increase rotation for controls but decreased rotation for the brain injured.

This paper describes two experiments. The first experiment demonstrates the existence of an interaction between instructions and brain injury; calling attention to tilt in the reproduced design benefits normal subjects more than brain injured. The second experiment

shows that this interaction effect and the effect due to reduced peripheral vision can be replicated, and that the two interactions can be combined to discriminate dull normals from brain injured.

EXPERIMENT 1: EFFECT OF NO-TILT INSTRUCTIONS

In all previous studies using the BDRT, the standard Wechsler-Bellevue Block Design instructions were used; i.e., the subject was not warned about rotation, he was simply told to reproduce the designs as accurately as possible. Thus, it was not possible to determine how much of the greater rotation of the brain injured was produced by their inattention to tilt.

On occasion, subjects were asked to correct the tilt in their completed designs. Some brain injured subjects were unable to do so, although they seemed to be trying. Control subjects generally had no difficulty when the tilt was called to their attention. This suggested that rotation was partly the result of inattention, but that the brain injured subjects, in addition, had difficulty perceiving rotation.

In Experiment 1, brain injured patients were compared with non-brain-injured controls under standard and no-tilt instructions. Four groups of 20 subjects were used: (a) brain injured with standard instructions (BS), (b) brain injured with no-tilt instructions (BN), (c) controls with standard instructions (CS), (d) controls with no-tilt instructions (CN).

Subjects

Forty male brain injured patients were selected from the Neurology and Neurosurgery Services at Walter Reed General Hospital. Table 1 shows the frequencies for the several types of injuries and a breakdown of these for approximate lateral localiza-

¹ In the BDRT the subject uses four blocks taken from the Wechsler-Bellevue Block Design subtest to reproduce blue and yellow designs, 1 inch square, painted on a white card, 6 inches square.

TABLE 1

CLASSIFICATION OF BRAIN INJURED SUBJECTS ACCORDING TO TYPE AND LOCATION OF INJURY: EXPERIMENT 1

Type of injury	Location of injury			N
	Left	Right	Bilateral	
Skull fracture	4	1	5	10
Gunshot wound	2	1		3
Closed head injury			4	4
Vascular disorders	4	4	2	10
Neoplasms	4	2		6
Encephalopathies, n.e.c.			7	7
Total	14	8	18	40

tion.² The category "encephalopathies, n.e.c" includes cases with encephalitis, Wilson's disease, and meningitis.

As can be seen in Table 1, the majority of the brain injured patients had relatively diffuse damage, difficult to localize. They were tested as soon after hospitalization as they were able to cooperate with the examiner, and understand instructions. At the time of testing, the length of hospitalization ranged from 1 to 7 months with a mean of 2.4 and a standard deviation of 1.6.

In a brief mental-status examination conducted prior to each test, the examiner judged 14 patients to be disoriented for time and/or place, with impaired memory. Patients were accepted for the study if they showed in practice trials that they understood the standard instructions for the Wechsler-Bellevue Block Design subtest. Prior to the occurrence of brain injury, all patients had been in general good health.

The 40 male controls were selected from the non-brain-injured, nonpsychiatric patient population at Walter Reed General Hospital. They had exhibited no signs of CNS damage on examination at hospital entry. A questionnaire was used to eliminate patients with a history of head injury.

The Army Classification Battery (ACB) (Montague, Williams, Lubin, & Giesecking, 1957), administered at Army entry previous to illness or injury, was available for 30 of the brain injured patients. Thirty of the controls were so selected as to match, individually, these 30 patients on the Pattern Analysis subtest of the ACB and on the time interval be-

tween first administration of the ACB and subsequent administration at hospital entry. Pattern Analysis was used because it appears to be a reliable, valid measure of spatial ability, relatively free from the effects of education. Matching on time since initial testing provides some control on age and rank. There were no significant differences in age among the four groups. The ages ranged from 18 to 50 years with a mean of 28.4 and a standard deviation of 8.3.

Procedure

The 40 target cards of the BDRT were placed, 1 by 1, in a single marked position on a table 32 inches wide by 50 inches long by 30 inches high. The surface of the table was painted a dull black. The target card was centered with respect to the length of the table and was 15 inches from the subject. The subject made his block designs within a circle of points, 6 inches in diameter, located at the table edge.

In the "no-tilt" groups (BN and CN) attention to correct orientation was induced by instructions, demonstrations of tilt, practice trials, and warnings when rotation occurred during the test. The remaining groups (BS and CS) received standard Wechsler Block-Design instructions.

When the subject indicated that he had completed a design, his reproduction and the target design were photographed with an overhead camera. Later degrees of rotation from the target design were measured from the film, using a ruler and protractor. Two individuals made independent measures of rotation for each target card, and adjusted their scores after discussion of major disagreements. The adjusted scores showed an average difference of about 2 degrees. The average of the two adjusted scores was used for each card.

Prior to administering the BDRT, all 80 subjects were given the Arithmetic, Vocabulary, and Block Design subtests of Wechsler-Bellevue. This Wechsler AVB combination was used to estimate intellectual level at the time of the study.

Results

Figure 1 shows the four groups in relation to the M^3 rotation score and the Wechsler AVB measure of intelligence.

The control group receiving the no-tilt instructions is significantly⁴ lower on rotation than the other three groups. If we remove this CN group, the remaining three groups do not differ significantly on M. Table 2 gives

³ R, the total degrees of rotation over the 40 cards, has a very skewed distribution. The logarithmic function $M = 100 [\log R - 1]$ was used to reduce skew and to reduce nonlinearity of regression on intelligence.

⁴ In this paper "significant" refers to the .05 level or better.

² Tables giving additional information on symptomatology, mental status, and special diagnostic studies for each patient are filed with the American Documentation Institute. Order Document No. 6871 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance \$1.75 for microfilm or \$2.50 for photocopies. Make checks payable to: Chief, Photoduplication Service, Library of Congress.

the mean and variance of *M* for each group of 20 subjects, as well as the correlation of *M* with AVB within each group. By the usual two-way analysis of variance of Table 2, the effect of instructions is significant at the .01 level, the brain injury effect is significant at the .05 level. The interaction of brain injury and instructions is significant at the .05 level, only if a one-tailed test is used.

The average *M* scores for the diagnostic classifications shown in Table 1 did not differ significantly, nor was there a significant laterality effect. There was no significant difference between average *M* scores for the groups judged to be oriented and disoriented in the mental status examination.

Discussion

When tilt is called to the subject's attention, the controls are able to reduce their rotation to about 4 degrees per card, quite close to the 2-degree average error measurement. However, the brain injured, even with no-tilt instructions, still average about 12 degrees of rotation per card. These results imply that most of the rotation by patients with recent general brain injury is due to impairment of perception rather than inattention.

In previous studies we were puzzled by the persistent negative correlation of about $-.50$ between rotation and intelligence. Clinical observation suggested that with standard block design instructions brighter subjects perceived the importance of correct orientation. Dull subjects seemed less concerned about the proper orientation of their design. If attention to rotation increases with intelligence, then the strength of the correlation between

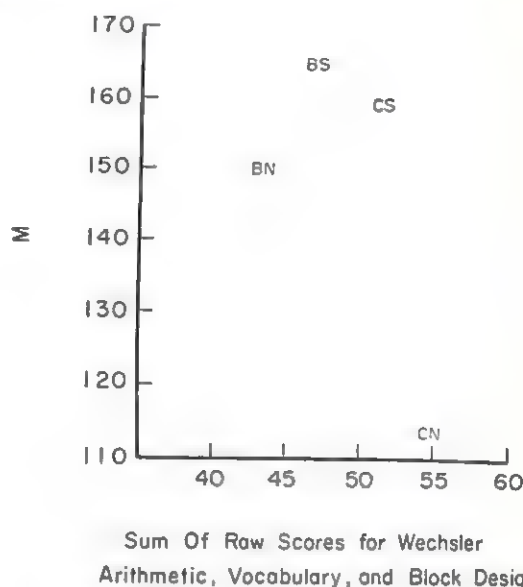


FIG. 1. Position of the four brain injury instruction groups with respect to rotation (*M*) and intelligence.

M and AVB should be reduced in the no-tilt instruction groups. As can be seen in Table 2, this is not true. The nature of the relation between intelligence and rotation remains a mystery.

EXPERIMENT 2: COMBINED EFFECT OF REDUCED PERIPHERAL VISION AND NO-TILT INSTRUCTIONS ON ROTATION

The purpose of this experiment is to replicate the two interactions found previously and to show that they may be combined to demonstrate a significant difference between dull normals and brain injured. The relation between intelligence and rotation is such that dull normals and brain injured rotate about

TABLE 2
MEANS, VARIANCES, AND CORRELATIONS WITH INTELLIGENCE, FOR THE ROTATION SCORE *M*, AS A FUNCTION OF BRAIN INJURY AND INSTRUCTIONS

Group	Rotation (<i>M</i>)		Correlation of <i>M</i> with AVB
	Mean	Variance	
Brain injured, standard instructions	165.15	2,054.13	$-.58$
Non-brain-injured, standard instructions	159.50	1,134.05	$-.57$
Brain injured, no-tilt instructions	150.40	2,008.78	$-.43$
Non-brain-injured, no-tilt instructions	114.35	558.13	$-.64$

Note.—The *N* in each group is 20.

TABLE 3
CLASSIFICATION OF BRAIN INJURED SUBJECTS ACCORDING TO TYPE AND LOCATION OF INJURY: EXPERIMENT 2

Type of injury	Location of injury			N
	Left	Right	Bilateral	
Skull fracture	2	2	4	8
Gunshot wound		1		1
Closed head injury			9	9
Vascular disorders	1			1
Encephalopathies, n.e.c.			1	1
Total	3	3	14	20

the same amount. But the interactions of peripheral vision and no-tilt instructions with brain injury are independent of intelligence. It follows that these two interactions could be combined to show a significant difference between dull normals and brain injured.

The reasoning is as follows: Suppose rotation scores are obtained under three conditions, (a) standard instructions, (b) standard instructions combined with reduced peripheral vision and (c) no-tilt instructions with unrestricted vision. Previous results indicate that brain injured and dull normals rotate equally often on Condition a. For the dull normals we would predict an increase in rotation from a to b, and a large decrease from b to c. For the brain injured patient there should be a decrease in rotation from a to b, and a slight drop from b to c. The difference score, $k = b - c$, should show a significantly higher mean for the dull normals since it adds the absolute value of the decreased rotation for no-tilt instructions to the increase in rotation due to reduced peripheral vision.

Subjects

Twenty male brain injured subjects were selected from the Neurology and Neurosurgery Services at Walter Reed General Hospital. Table 3 gives frequencies for the various diagnoses. There are proportionally more bilateral cases, but in other respects the group resembles the brain injured subjects of Experiment 1. At the time of testing the length of hospitalization ranged from 1 to 9 months, with a mean of 2.9 and a standard deviation of 2.2.

Twenty control patients were selected so as to match each brain injured subject on the Pattern Analysis score at time of Army entry, and on time

between first and second administration of the ACB. These controls were selected from the same population described for Experiment 1. A dull normal group was formed by selecting 20 control patients who had made a score of 80 or below on Pattern Analysis at the time of Army entry. (A score of 80 is one standard deviation below the mean.) The mean age for the three groups was 26.9, the standard deviation 8.6, and the ages ranged from 17 to 50 years. There was no significant differences between the means of the three groups.

Procedure

Each subject was asked to designate his preferred eye, and monocular vision was used throughout. Let A designate the first 20 trials of the BDRT. Let B designate the second 20 trials of the BDRT. C represents 20 additional trials obtained by a clockwise, 90° rotation of each of the first 20 BDRT cards. Every subject was tested in the same way: first on A with standard instructions, then on B with Shapiro's field reducer,⁵ finally on C with no-tilt instructions and unrestricted monocular vision.

Results

Figure 2 shows the average degrees of rotation per design. As predicted for B, the field reducer condition, rotation increases for the dull normals and normals, whereas the brain injured show a decrease in rotation. All three groups show decreased rotation under no-tilt instructions, but as expected, the brain injured show the smallest improvement.

⁵ The field reducer is a mask fashioned from a table tennis ball which covers the eye. It permits central vision through a hole about 6 millimeters in diameter.

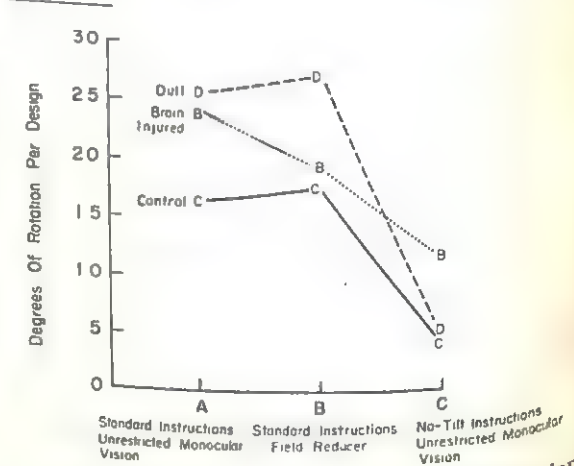


FIG. 2. The effect of restricted peripheral vision and no-tilt instructions on rotation for controls, dull normals, and brain injured.

TABLE 4

CHANGES IN DEGREES OF ROTATION PER CARD DUE TO
REDUCED PERIPHERAL VISION AND NO-TILT
INSTRUCTIONS

Group	Mean		
	A	B - A	K = B - C
Normals	16.578	1.138	13.427
Dull normals	25.460	1.768	22.037
Brain injured	23.958	-4.362	7.642
	Variance		
	A	B - A	B - C
Normals	98.485	83.933	156.130
Dull normals	254.401	327.775	480.944
Brain injured	135.052	60.949	206.615

Note.—The *N* in each group is 20.

Table 4 gives the basic data for estimating the effect of each condition. Under Condition A, with standard instructions and unrestricted monocular vision, the brain injured and dull normals have about the same amount of rotation. As expected, the normals have a significantly lower mean than the other two groups.

The column labeled B-A measures the effect of reducing peripheral vision. Both the normals and the dull normals show a slight increase in rotation, averaging about 1 or 2 degrees per card, whereas the brain injured have a significant decrease in rotation.

The column labeled $K = B - C$ is equal, algebraically, to $(B-A) - (C-A)$ and therefore is equivalent to adding the effect of reduced peripheral vision and subtracting the effect of no-tilt instructions. The dull normals are significantly higher than the brain injured on this combined measure of interaction.

K has a statistically significant correlation of .37 with the dichotomous criterion, brain injured vs. dull normals. The multiple regression of the dichotomous criterion on the scores A, B, and C gave a multiple correlation of .42. This does not differ significantly from the .37 validity of K . In other words, the *a priori* function, $K = B - C$, is as good as the best empirical discriminating function. Neither function, however, is very useful for diagnostic purposes. The best (i.e., maximum

likelihood) cutoff point yields only about 60% correct classification. The average rotation scores for the diagnostic groups shown in Table 3 do not differ significantly, nor was there a significant difference between the averages for disoriented and oriented patients.

Discussion

Rotation in brain injured and normals occurs intermittently much as would be anticipated from sporadic spells of inattention. Directing the attention of normal subjects to tilt does reduce rotation to an amount close to the error of measurement, but brain injured patients improve only slightly. The paradoxical finding that reducing peripheral cues causes normals to rotate more, but actually improves the performance of the brain injured subject implies that relevant, peripheral cues may cause orientation error in patients with brain injury. It may be inferred that there is a malfunctioning of the general integrating mechanism in the brain injured subject, such that relevant peripheral cues hamper performance by producing distorted perception.

M. B. Shapiro (1952) hypothesizes that the greater rotation for brain injured subjects is due to an increase in cortical inhibition caused by trauma. Thus, the brain injured subject is rendered peripherally blind, and integration fails because the peripheral cues are not transmitted by the cortex. Therefore, Shapiro's prediction would be that the field-reducer would have no effect on rotation for the brain injured subjects. The data of this and the previous experiment (Williams et al., 1956) indicate, however, that the field-reducer facilitates correct orientation by the brain injured.

The results obtained by Strauss and Lehtinen (1947) appear to be similar to ours. They state that brain injured subjects are easily distracted by stimuli; therefore reducing the display to its essentials will improve performance. The field-reducer used in the present experiment may prevent peripheral stimuli from distracting the brain injured subject, thus enabling him to concentrate more effectively on the target. In place of Shapiro's "inattention" hypothesis, we would substitute the notion that for the brain injured, relevant peripheral cues provide con-

fusing and distracting information about the visual frame of reference.

SUMMARY

Experiments were conducted to confirm the existence of two interactions of brain injury and experimental conditions on block design rotation: (a) Instruction to pay attention to the tilt of the reproduced designs resulted in a greater decrease of rotation for both normal and dull normal controls than for the brain injured. (b) Restricting peripheral vision increased rotation for normal and dull normal controls, but decreased it for the brain injured. Although the difference in patterns of performance for dull normals and brain injured was statistically significant, it was not great enough to furnish a basis for individual diagnosis.

The results from this and previous experiments imply that the basic difficulty for brain injured subjects is not a failure of attention or peripheral blindness, but is a generalized defect of integration such that relevant peripheral cues cause perceptual distortion.

REFERENCES

- BENDER, M. B., & TEUBER, H. L. Spatial organization of visual perception following injury to the brain. *Arch. Neurol. Psychiat.*, 1948, 59, 39-62.
- GOLDSTEIN, K., & SCHEERER, M. Abstract and concrete behavior. *Psychol. Monogr.*, 1941, 53(2, Whole No. 239).
- HANVÍK, L. J., & ANDERSON, A. L. The effect of focal brain lesions on recall and on the production of rotations in the Bender-Gestalt test. *J. consult. Psychol.*, 1950, 14, 197-198.
- MONTAGUE, E. K., WILLIAMS, H. L., LUBIN, A., & GIESEKING, C. F. The use of Army tests for the assessment of intellectual deficit. *U. S. Armed Forces med. J.*, 1957, 8, 883-892.
- PASCAL, G. R., & SUTTELL, BARBARA J. *The Bender-Gestalt test*. New York: Grune & Stratton, 1951.
- SHAPIRO, M. B. Experimental studies of a perceptual anomaly: I. Initial experiments. *J. ment. Sci.*, 1951, 97, 90-110.
- SHAPIRO, M. B. Experimental studies of a perceptual anomaly: II. Confirmatory and explanatory experiments. *J. ment. Sci.*, 1952, 98, 605-617.
- SHAPIRO, M. B. Experimental studies of a perceptual anomaly: III. The testing of an explanatory theory. *J. ment. Sci.*, 1953, 99, 394-409.
- STRAUSS, A. A., & LEHTINEN, LAURA E. *Psychopathology and education of the brain-injured child*. New York: Grune & Stratton, 1947.
- WILLIAMS, H. L., LUBIN, A., GIESEKING, C. F., & RUBINSTEIN, I. The relation of brain injury and visual perception to block design rotation. *J. consult. Psychol.*, 1956, 20, 275-280.
- YATES, A. J. An experimental study of the block design rotation effect with special reference to brain damage. Unpublished doctoral dissertation, University of London, 1954.

(Received October 26, 1960)

A NOTE ON TIME OF FIRST RESPONSES IN RORSCHACH PROTOCOLS

ALVIN G. BURSTEIN¹

University of Michigan

It is conventional in preparing the summary of a subject's Rorschach performance for diagnostic purposes to compute the mean time of first response (T/1R) for the series of 10 cards. The object is to obtain a measure of central tendency—to estimate the "typical" T/1R for that subject—(a) so that a subject's typical T/1R can be compared with typical times for nosological groups and (b) so that the T/1R on a specific card for a particular subject can be compared with that same subject's typical T/1R. Examples of the clinical value of such comparisons have been furnished by Beck (1949). Depressed and organically deteriorated patients have a typically slower T/1R than do normals (a comparison of the first type mentioned), while neurotic shock can often be identified in terms of a subject's departure from his own typical T/1R (a comparison of the second type).

Although the general practice is to compute the mean T/1R, it might be well to consider whether the median might not be more appropriate. Since the population of response times can extend no lower than zero seconds but up to very high values, and since most response times are clustered near the low end, the population is skewed, and the mean and the median will not coincide. A choice between these two measures of central tendency could be based on the same arguments that favor the use of the median over the mean in de-

scribing the "average" American's income; the sensitivity of the mean to extreme values makes it appear preferable to have that figure below which half the cases fall and above which half the cases fall, that is, the median.

The kind of distortion to which the mean T/1R may be subject is illustrated in a case reported by Beck (1949, pp. 281-287). In evaluating evidence for neurotic shock on several cards, Beck used, as a basis of comparison not the overall mean T/1R of 65.1 seconds, but rather a corrected mean T/1R—28 seconds—obtained by ignoring the three largest values. It should be noted that, because it is less sensitive to extreme values, the median T/1R—33 seconds—could have been used without such correction. Since it is difficult in such cases to make the subjective judgment of which and how many extreme values to drop, the use of the median would appear advantageous.

In an effort to supply some normative information, the median T/1R was computed for 154 of the 157 protocols collected as a normative sample by Beck, Rabin, Theissen, Molish, and Thetford (1950; the remaining three were not available at the time of the analysis). These protocols had a mean median T/1R of 25.6 seconds as compared with a mean mean T/1R of 32.5 seconds.

The really critical issue in deciding which measure of central tendency to employ is what we wish to represent by that measure. It is characteristic of the median that it will represent the typical response time in the sense that exactly half of the subject's response time will be shorter than the median

¹ The author is indebted to S. J. Beck of the University of Chicago and to Sheldon Korchin of Michael Reese Hospital for their assistance in obtaining access to the normal Rorschach protocols.

and half longer. As we have seen, the sensitivity of the mean to extreme values can give us a typical time much higher than this. It is suggested therefore that the median more adequately represents the typical response time, and substituting the median for the mean in Rorschach protocols should help make the clinical use of the time of first response more meaningful.

REFERENCES

- BECK, S. J. *Rorschach's test*. Vol. 2. New York: Grune & Stratton, 1949.
- BECK, S. J., RABIN, A. I., THEISSEN, W. G., MOLISH, H. B., & THETFORD, W. N. The normal personality as projected in the Rorschach test. *J. Psychol.*, 1950, 30, 241-298.

(Received September 28, 1960)

EGO STRENGTH AND CONFLICT DISCRIMINATION: A FAILURE OF REPLICATION

JACK BLOCK

University of California, Berkeley

Korman (1960) recently has reported a study wherein the latency of psychophysical judgments was found to be related to scores on Barron's *Es* (ego strength) scale. Subjects scoring low on the *Es* scale were found to discriminate more slowly than subjects scoring high. Further, as the objective difficulty of decision was increased, this difference in latency of judgment was found to increase. These results were sought specifically, as one route to the construct validation of the *Es* scale. The general principle of validating a measure by relating it to a very different index of the underlying construct is of course a worthy one and it is with reluctance that the present note introduces data which fail to confirm the Korman finding.

In a previous study (Block & Petersen, 1955) after which the Korman experiment in part was patterned, latency scores for both easy and difficult judgment-discriminations were available which are fully equivalent to the latency scores employed by Korman. Also available were scores for each of the 53 subjects on the *Es* scale. For the easy decision situation and separately for the difficult decision situation, the 53 subjects were ordered with respect to their decision latencies. The *Es* scores of the fastest 25 subjects were then contrasted with the *Es* scores of the slowest 25 subjects (the intermediate three subjects being omitted for reason of computational convenience). The fast deciders in an objectively easy decision situation had a mean *Es* score of 50.96 with an *SD* of 4.02; slow deciders had a mean of 50.88 with an *SD* of 5.03. The fast deciders in an objectively ambiguous situation had a mean *Es* score of 50.72 with an *SD* of 4.22; the slow deciders had a mean *Es* score of 51.16 with an *SD* of 4.88. Obvi-

ously, in this study there is no relation between *Es* scores and the ability to rapidly resolve discrimination conflicts.

How may such an empirical discrepancy be understood? What factors may be contributing to a finding of relationship in the one study and the absence of a relationship in the other?

One immediately obvious consideration is that the samples employed in the two studies are radically different. The Korman study used 47 psychiatric inpatients, all presumably with sufficient internal and manifest psychopathology to warrant commitment. The Block-Petersen study employed Air Force captains, all presumably individuals within the normal range of adjustment. The mean *Es* score for the Korman sample was about 41, well below the Block-Petersen sample mean of 51.32. The *Es* standard deviation of the Korman sample is 6.75, somewhat but not significantly higher than the *SD* of 4.54 in the Block-Petersen sample. These are important differences for the psychological significance of a given *Es* score in the one sample may not correspond to its meaning in the other sample. Simply at the quantitative level, the differences in the *Es* means of the two samples suggest that the high *Es* scorers of the Korman sample were relatively low scorers when considered relative to the Block-Petersen sample.

It would be presumptuous to discuss the comparative merits of these two samples for an appropriate test of the Korman hypothesis. Properly, many more samples should be studied so that a pattern of results and their converging implication may appear. It would seem clear, however, in this instance and in many more that doubtless could be recounted,

that the characteristics of the sample being studied must be recognized explicitly as modifying in decisive ways the relationships observed (Block, 1955). To the plea of Campbell and Fiske (1959) for "heterotrait" and "heteromethod" validity must be added the requirement of *heterogroup* validity as well.

REFERENCES

- BLOCK, J. The difference between *Q* and *R*. *Psychol. Rev.*, 1955, 62, 356-358.
- BLOCK, J., & PETERSEN, P. Some personality correlates of confidence, caution, and speed in a decision situation. *J. abnorm. soc. Psychol.*, 1955, 51, 34-41.
- CAMPBELL, D. T., & FISKE, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.*, 1959, 56, 81-105.
- KORMAN, M. Ego strength and conflict discrimination: An experimental construct validation of the ego strength scale. *J. consult. Psychol.*, 1960, 24, 294-298.

(Received October 7, 1960)

BRIEF REPORTS

ANALYSIS OF THE WISC PERFORMANCE OF BRAIN DAMAGED AND EMOTIONALLY DISTURBED CHILDREN¹

VINTON N. ROWLEY

State University of Iowa

A diagnostic question, to which the clinical psychologist is often called upon to contribute, is the differentiation of emotionally disturbed and brain damaged children with average or superior intelligence. In this situation, it is not uncommon to employ the discrepancy between the Verbal scale and the Performance scale of the WISC and similar indices as aids in the identification of brain damage. However, there is little empirical evidence to support these particular uses of the WISC as a diagnostic technique.

The present investigation analyzed the WISC performances of emotionally disturbed and brain damaged children with respect to such characteristics as overall level and pattern of performance, differences between Verbal and Performance scale quotients, and differences between subtest means. The emotionally disturbed group consisted of 30 children (mean age = 10.5 years) who had been seen in the Child Psychiatry Serv-

ice, University of Iowa Psychopathic Hospital, because of behavioral maladjustment and who had been diagnosed as emotionally disturbed with no evidence or history of cerebral disease. The brain damaged group consisted of 30 children (mean age = 10.6 years) who had been seen in the Pediatrics Clinic, University Hospitals, and who showed unequivocal evidence of disease involving one or both cerebral hemispheres.

In order to provide for as precise a comparison as possible, certain restrictions were observed in the selection of subjects. The subjects were individually matched with respect to sex, CA, and WISC Full Scale IQ to minimize the effects of these variables on performance patterns. A minimal IQ of 83 was established in order to exclude defective children.

The essential findings were: (a) there was no significant difference between the two groups with respect to either Verbal scale or Performance scale IQ; (b) Verbal scale-Performance scale relationships were not significantly different in the two groups; (c) the profiles of subtest scores in the two groups were not significantly different; (d) none of the individual subtest scores showed a significant intergroup difference.

The general conclusion to be drawn from these findings is that, when overall level of performance is controlled, examination of the pattern of WISC performance does not provide a basis for differentiating nondefective brain damaged children from nondefective emotionally disturbed children.

(Received January 3, 1961)

¹ This investigation was supported by a grant (B-616) from the National Institute of Neurological Diseases and Blindness, United States Public Health Service. The writer is indebted to Arthur L. Benton for aid in planning and executing the study.

An extended report of this study may be obtained without charge from Vinton N. Rowley (Department of Psychiatry, State University of Iowa; Iowa City, Iowa) or for a fee from the American Documentation Institute. Order Document No. 6872 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance \$1.75 for microfilm or \$2.50 for photocopies. Make checks payable to: Chief, Photoduplication Service, Library of Congress.

SOCIAL DESIRABILITY IN THE RATINGS OF INVOLVED AND NEUTRAL JUDGES¹

GEORGE LEVINGER
Western Reserve University

Research on personality rating devices has revealed a high positive correlation between probability of item endorsement and the item's perceived social desirability (e.g., Edwards, 1953, 1957b). Recent studies have been mainly concerned with ascertaining the pervasiveness of this correlation, and with constructing instruments which would reduce this bias. The present report, dealing with one aspect of the problem from a somewhat different perspective, conceives of personality ratings as the reflection of "true score" and error displacement.

On the one hand, it is hypothesized that desirable traits are truly more common than undesirable ones. It is also hypothesized that raters will distort their ratings in a desirable direction, to the extent of their attraction toward the object rated.

Data are based on a study of 31 family triads, consisting of a father, mother, and 11-year-old child. As part of a larger study, check list ratings of family members were obtained from the parents and from either clinicians acquainted with the family or from teachers acquainted with the child.

Regarding the first hypothesis, it was found that the item endorsement frequencies of all raters for all objects correlated positively with the *SD* scale values of the items (Edwards, 1957a), though not necessarily at a statistically

significant level. The correlations ranged from .08, for the clinicians' ratings of disturbed children, to .83, for the school parents' descriptions of themselves. Such a finding is not surprising, considering the lengthy socialization process in any human culture. It seems that an objective judge would tend to place almost any person above the neutral point of social desirability.

Regarding the second hypothesis, parents' ratings of their children—and of themselves—were consistently more favorable than those of the teachers or clinicians. The findings, while limited to the correlational data mentioned above, tend to support the hypothesis.

The findings are not in themselves novel. Yet their implication is that investigations of social desirability loadings should not limit their focus to item content alone, but also concern themselves with the nature of the judge-object relationship.

For example, a disturbed person will probably describe himself less favorably than will a non-disturbed one. It would seem that this is not so much due to the former's different interpretation of item content as to his unfavorable state of personal self-attraction. And, when one finds positive changes in self-description among patients in psychotherapy, the scores may merely reflect changing attraction between subject and object.

REFERENCES

- ¹ An extended report of this study may be obtained without charge from George Levinger (Western Reserve University; Cleveland, Ohio) or for a fee from the American Documentation Institute. Order Document No. 6873 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance \$1.25 for microfilm or \$1.25 for photocopies. Make checks payable to: Chief, Photoduplication Service, Library of Congress.
- EDWARDS, A. L. The relationship between the judged desirability of a trait and the probability that the trait will be endorsed. *J. appl. Psychol.*, 1953, 37, 90-93.
- EDWARDS, A. L. Social desirability and probability of endorsement of items in the interpersonal check list. *J. abnorm. soc. Psychol.*, 1957, 55, 394-396. (a)
- EDWARDS, A. L. *The social desirability variable in personality assessment and research*. New York: Dryden, 1957. (b)

(Received January 10, 1961)

MEASUREMENT OF THE SEVERITY OF DISORDER IN SCHIZOPHRENIA BY MEANS OF THE HOLTZMAN INKBLOT TEST¹

RICHARD A. STEFFY AND WESLEY C. BECKER

University of Illinois

Elgin Prognostic Scale ratings of behavioral and case history data were correlated with genetic level ratings (Becker, 1956) of Holtzman Inkblot responses on a sample of 36 Veterans Administration, hospitalized schizophrenics. A product-moment r of $-.36$ ($p < .05$) supported the prediction that subjects with poorer Elgin Prognostic Scale scores tend to give more diffuse, undifferentiated, immature responses to inkblot stimuli. Although the relationship between genetic level and Elgin ratings was not as high as the one found by Becker (1956) using the Rorschach, differences between samples in duration of present hospitalization were shown to attenuate the relationship. Longer hospitalization leads to lower Elgin ratings on some scales (duration of psychosis, social withdrawal) and to improved functioning on inkblot tests (as precipitating stresses

are removed). Partialing out the effect of duration of hospitalization increased the correlation between the Elgin scale and the genetic level scoring of the Holtzman to $-.46$.

Additional analyses of the Holtzman test were made to explore the limits of its potential in this area. Although based on a small sample and needing cross-validation, an item analysis revealed a best subset of 13 Holtzman cards that correlated $-.53$ (and $-.64$ after partialing out duration of hospitalization) with the Elgin scale criterion ratings. The best linear combination of five Holtzman variables entering into the genetic level scoring system did nearly as well in predicting the Elgin scale as the pattern scores of the genetic level scoring system. It is concluded that more extensive studies of this type with Holtzman test offer promise of producing good measures of degree of pathology in schizophrenia.

REFERENCE

- BECKER, W. C. A genetic approach to the interpretation and evaluation of the process-reactive distinction in schizophrenia. *J. abnorm. soc. Psychol.*, 1956, 53, 229-236.

(Received January 20, 1961)

¹ An extended report of this study may be obtained without charge from W. C. Becker (Department of Psychology, University of Illinois; Urbana, Illinois) or for a fee from the American Documentation Institute. Order Document No. 6874 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance \$1.25 for microfilm or \$1.25 photocopies. Make checks payable to: Chief, Photoduplication Service, Library of Congress.

LYONS RELATIONSHIP SCALES: A STUDY OF RELIABILITY¹

ELIZABETH L. GOUCHER,² LAURA E. RIGGS, HERMAN Y. EFRON,
REBECCA F. MYERS, AND EMILY R. SCANLAN
Veterans Administration Hospital, Lyons, New Jersey

An important area of interpersonal functioning is the relationship between an individual and the family members with whom he is living. In conjunction with a study of the effectiveness of casework with relatives of hospitalized psychiatric patients a review of the literature failed to reveal a scale appropriate for assessing, or measuring changes in, the patient-family relationship. This study reports the development of such scales and an evaluation of their reliability.

The Lyons Relationship Scales consist of two parts: Schedule I, The Relationship from the Viewpoint of the Patient; and Schedule II, The Relationship from the Viewpoint of the Relative. They were designed to be used after an interview with the patient and an interview with the relative. Each schedule consists of 13 items reflecting the description given by the patient of his relative (or vice versa) with references to specific behavioral characteristics we thought of as being influential in determining the nature of a relationship between close relatives. Specifically the items dealt with demandingness, consideration of mate's views and feeling, interchange of ideas, adaptability, discussion of problems, sharing of work, degree of overprotectiveness, expression of affection, degree of hostility, money arrangements, and consideration of patient's illness. In addition there is an item on each schedule concerning each relative's global evaluation of the "goodness" of the relationship and one item which calls for the rater's evaluation of the relationship based on an overall appraisal of the elicited material.

¹ An extended report of this study may be obtained without charge from Emily Scanlan (Chief, Social Work Service, Veterans Administration Hospital; Lyons, New Jersey) or for a fee from the American Documentation Institute. Order Document No. 6871 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance \$1.75 for microfilm or \$2.50 for photocopies. Make checks payable to: Chief, Photoduplication Service, Library of Congress.

² Now at the New Jersey Neuropsychiatric Institute, Princeton, New Jersey.

The reliability of these instruments was assessed by studying results obtained from Schedule I, The Relationship from the Viewpoint of the Patient. The assumption was made that since the items on Schedule II were analogous to those on Schedule I, ratings based upon interviewing a relative would be at least equally as reliable as those based upon interviewing a patient. Thirty-four patients were interviewed and independently rated by a panel of four social workers.

The index Q (Hester, 1957 unpublished) was utilized to assess the reliability of the scale. Q differs from most indices of correlation in that it takes into account not only the degree to which different raters agree on an item, but also the extent to which the item differentiates one respondent from another. Only two of the items (sharing of work and consideration of patient's illness) did not attain a satisfactory Q value. The item on adaptability was discarded since it was too often considered unrateable. Two others (discussion of problems and money arrangements) were borderline and were tentatively retained pending further study.

As a further attempt to assess the reliability of the scale the statistic κ (kappa) developed by Cohen (1960) was utilized. One of the major respects in which this statistic differs from Q is that rather than being based upon dichotomizing each variable, all the item rating points are used. Analysis revealed moderate agreement for all the scale items except for the two that did not attain a satisfactory Q .

We feel that these scales, with face validity and demonstrated reliability, may be of value in that they constitute a measure of an important area of interpersonal relationships.

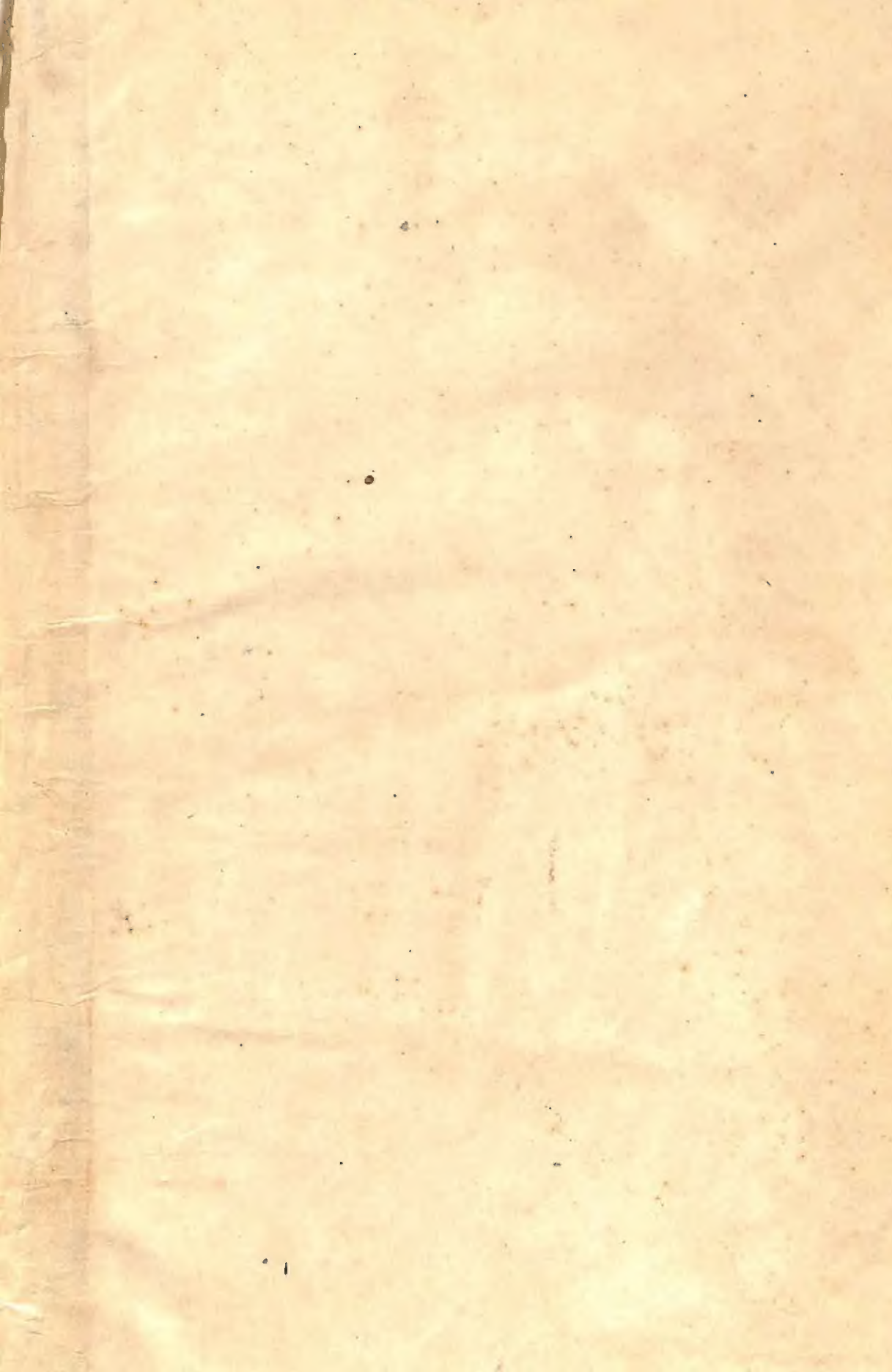
REFERENCES

- COHEN, J. A coefficient of agreement for nominal scales. *Educ. psychol. Measmt.*, 1960, 20, 37-46.
HESTER, R. Mathematical treatment of data. Unpublished manuscript, Psychiatric Evaluation Project, Veterans Administration Hospital, Washington, D. C., 1957.

(Received February 20, 1961)









21 JAN 1966

